Parameter-free Molecular Super-Structures Quantification in Single-Molecule Localisation Microscopy

Parameter-free Molecular Super-Structures Quantification in Single-Molecule Localisation Microscopy

Mattia Marenda[1,2,*], Elena Lazarova[1], Sebastian van de Linde[3], Nick Gilbert[1,*] and Davide Michieletto[1,2]

[1] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Crewe Road, University of Edinburgh, Edinburgh EH4 2XU, UK

[2] SUPA, School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Road, Edinburgh EH9 3FD, UK

[3] SUPA, Department of Physics, University of Strathclyde, Glasgow G4 0NG, UK

*Correspondence to Mattia Marenda: mattia.marenda@igmm.ed.ac.uk; Nick Gilbert: nick.gilbert@ed.ac.uk

eTOC

Marenda et al. introduce a parameter-free algorithm to quantify super-structures and connected clusters in SMLM datasets. The algorithm is tested on simulated and experimental datasets demonstrating that it can be used as an unbiased tool to extract information beyond simple clustering.

Abstract

Understanding biological function requires the identification and characterisation of complex patterns of molecules. Single-Molecule Localisation Microscopy (SMLM) can quantitatively measure molecular components and interactions at resolutions far beyond the diffraction limit, but this information is only useful if these patterns can be quantified and interpreted. We provide a new approach for the analysis of SMLM data that develops the concept of structures and super-structures formed by inter-connected elements, such as smaller protein clusters. Using a formal framework and a parameter-free algorithm, (super-)structures formed from smaller components are found to be abundant in classes of nuclear proteins, such as heterogeneous ribonucleoprotein particles (hnRNPs), but are absent from ceramides located in the plasma membrane. We suggest that mesoscopic structures formed by interconnected protein clusters are common within the nucleus and have an important role in the organisation and function of the genome. Our algorithm, “SuperStructure”, can be used to analyse and explore complex SMLM data and extract functionally relevant information.

Introduction

Single-molecule localisation microscopy (also known as SMLM) (van de Linde et al., 2011; Schermelleh et al., 2010; Henriques et al., 2011; Sauer and Heilemann, 2017) is now commonly employed for quantitative analysis of molecular structures and interactions both in cell-based (Cisse et al., 2013; Kapanidis et al., 2018; Chong et al., 2018) and *in vitro* experiments (Revyakin et al., 2006; Deniz et al., 2008). Unlike other light microscopy techniques, SMLM achieves resolutions far beyond the diffraction limit and its typical output is a list of 3D coordinates (or localisation events) that are naturally analysed using efficient clustering algorithms borrowed from quantitative big-data analysis and even astronomy (Owen et al., 2010; Sengupta et al., 2011; Garcia-Parajo et al., 2014; Baumgart et al., 2016; Spahn et al., 2016; Griffié et al., 2016). However, traditional clustering algorithms rely on user-defined parameters that are intrinsically intertwined with the notion of similarity that is necessary to define a cluster. These parameters can be either hypothesised by physical intuition or inferred via pre-emptive analysis (Burgert et al., 2017; Williamson et al., 2020; Malkusch and Heilemann, 2016), yet their choice has a significant impact on the results, in turn hindering the portability of clustering algorithms and the comparison between different datasets.

At the same time, recent evidence suggest that assemblies of proteins (Brangwynne et al., 2015; Larson et al., 2017; Strom et al., 2017; Sabari et al., 2018; Cho et al., 2018; Maharana et al., 2018; Chong et al., 2018) and chromatin (Bintu et al., 2018; Boettiger et al., 2016; Frank and Rippe, 2020) form functional complex structures that are not fully captured by standard clustering algorithms. For example, the hnRNP protein SAF-A is suggested to form a dynamic and functional mesh-like structure while interacting with RNA to maintain transcriptionally active genomic loci in a decompacted configuration (Nozawa et al., 2017; Michieletto and Gilbert, 2019). Other examples include SC35, a nuclear protein involved in RNA splicing and chromatin elongation (Lin et al., 2008) and that displays localised nuclear speckles (Xie et al., 2006; Jackson et al., 2000), or actin and microtubules which form elongated and inter-connected networks involved in cell motility and division, as well as in the synaptic plasticity of dendritic spines (Resch et al., 2002; Rogers et al., 2003; Izeddin et al., 2011). Additionally, recent super-resolution studies indicate that chromatin is also functionally organised in connected nano-scale compartments (Prakash et al., 2015; Szabo et al., 2018; Nir et al., 2018; Maiser et al., 2020). Rapidly evolving methods of chromatin tracing (Boettiger et al., 2016; Wang et al., 2016; Beliveau et al., 2015; Nir et al., 2018; Bintu et al., 2018) and super-resolved imaging of the accessible genome (Xie et al., 2020) require sophisticated algorithms to analyse the topology of the generated paths (Goundaroulis et al., 2019). In order to understand the relationship between these complex structures and the underlying biological mechanism and functions of the genome (Bronshtein et al., 2015; Khanna et al., 2019; Leidescher et al., 2020 *Preprint*; Smeets et al., 2014) a more sophisticated and standardised analysis of SMLM data is urgently required.

It is clear that quantification of complex structures is a ubiquitous problem in molecular and cell biology and it is intimately connected to cellular function. Motivated by this problem, here we introduce a new algorithm termed "SuperStructure", which extends in a novel and original way the popular density-based clustering algorithm DBSCAN. SuperStructure allows (i) a parameter-free detection and quantification of complex structures made of connected clusters in SMLM data and (ii) a parameter-free quantification of the density of molecules within clusters.

Here, we demonstrate the capabilities of SuperStructure on simulated datasets and then use it to analyse two groups of experimental datasets: (i) nuclear proteins involved in RNA processing, namely SAF-A, hnRNP-C and SC35 and (ii) ceramides lipids involved in cellular trafficking at the membrane. We find that interconnections between clusters are abundant in classes of proteins in the hnRNP family and that they are surprisingly absent from ceramides, suggesting this feature is relevant for the biological function of SAF-A and hnRNP-C. Therefore, SuperStructure enables us to discover new facets of protein organisation in human cells and provides a better understanding of the molecular mechanisms underlying the organisation of sub-cellular (super-)structures.

Finally, since SuperStructure is parameter-free, it provides the community with a standardised tool for the discovery and quantification of complex patterns in SMLM data. Furthermore, beyond helping our understanding of complex biological structures, it might be used to assess the fluorophore blinking quality and thus offers versatility in assessing also technical imaging properties (van de Linde and Sauer, 2014; Hennig et al., 2015; Siegberg and Herten, 2011).

Results

Super Structure Algorithm

SuperStructure is best explained in relation to the well-known DBSCAN algorithm. DBSCAN detects clusters by grouping together high-density localisations and classifies as outliers low-density ones (Ester et al., 1996). In practice, DBSCAN determines that a localisation is part of a cluster if more than N_{min} other localisations are found within a neighbourhood distance ε (or if it is part of the neighbourhood of another localisation with this property). Conversely, SuperStructure extracts connectivity information from the rate at which the number of detected clusters N_c changes with the neighbourhood radius ε for a fixed N_{min} (see Fig.1). Indeed, the curves $N_c(\varepsilon)$ contain important overlooked information about the structure of connections. To simplify the analysis, and without loss of generality, we set $N_{min} = 0$, which means that we do not require a minimum number of localisations within the neighbourhood to define a cluster. As a consequence, $N_c(\varepsilon)$ is necessarily a monotonically decreasing function as for $\varepsilon = 0$ every localisation is detected as a single cluster and increasing ε yields fewer but larger clusters. Following on, the rate at which N_c decays with ε is an indicator of how quickly localisations, and then clusters of localisations, coalesce, thus indicating how much localisations and clusters are connected.

The $N_c(\varepsilon)$ curves provided by SuperStructure identify different clustering regimes (Fig.1): the first (small ε) regime describes the merging of localisations within clusters (intra-cluster regime); the second (intermediate ε) regime captures the growth of clusters into super-structures (first super-cluster regime) and finally the third (large ε) regime describes the merging of super-clusters into higher-order super-structures (second/third super-cluster regimes). The $N_c(\varepsilon)$ curve in the first regime typically follows a Poissonian function (Eq.1) and its decay rate is related to the density of emitters ρ_{em} within the clusters (see Methods and Figs.1 and S1). The width of the Poisson function also sets the critical value of ε at which this first regime is expected to end (Eq.2). On the other hand, the decay in the second and third regimes follows

an exponential decay with characteristic length-scale λ and are highly dependent on the connectivity between (super-)clusters, as well as on the density of (super-)clusters (Eq.4).

The number of super-cluster regimes depends on the homogeneity of both cluster distribution and connections. In the two extreme cases of a completely connected or unconnected homogeneous distribution of clusters, we expect a single super-cluster regime. However, while in the former case this regime is exponential (because the clusters are connected), in the latter it assumes a Poissonian functional form (see respectively Eqs.4 and 3). This is not surprising, as free (unconnected) clusters that are randomly distributed behave (on a larger scale) as single emitters inside clusters (see Methods and Fig.S1). Also, in the case of clusters embedded in a random distribution of other localisations (such as noise), we obtain a Poissonian decay. Importantly, a random distribution of localisations (also at high density) is different from “connected” clusters, where nearby localisations are mostly distributed in between clusters. As a result, the curves generated by SuperStructure allow us to identify the presence/absence of connectivity by investigating the functional form of the curves, as well as to extract their decay rates.

In heterogeneous systems that display a mix of randomly dispersed localisations/clusters and connected ones over similar length-scales, we strongly recommend restricting the analysis with ROIs over sub-regions that display qualitatively similar phenotypes. A good example of heterogeneous system is given by the nuclear protein SC35, which we analyse below. Restricting the analysis to ROIs is also recommended when quantifying nuclear or cellular sub-structures that display boundaries. Masking localisations falling outside these boundaries allows SuperStructure to generate cleaner curves that are easier to interpret.

In order to quantify the intra-cluster density and (super-)cluster connectivities, one needs to define boundaries between regimes and to fit every regime with the corresponding function (see Eqs.1, 3 and 4). Regime boundaries and fitting ranges can be either selected manually (where curves change their decay properties) or by rigorously running a pre-emptive goodness-of-fit test. For instance, once the rough regime range has been identified and fitted, one can modify the fit window to identify the boundaries of the regime outside which the fit is no longer acceptable. Arguably, the optimum regime is found by identifying the best goodness-of-fit window (e.g. the range with the minimum chi squared). It is also possible to define a single function fitting the entire curve by (a) defining a piecewise function where every "piece" is the fit of the corresponding regime or by (b) adding together the contribution of the different regimes (appropriately weighted).

The work-flow for the application of SuperStructure is shown in Fig.1 and is described in detail in Methods. Additionally, the codes and scripts are open source and available at git repository (see below).

Characterising SuperStructure Feature Extraction from Simulated SMLM data

To evaluate the performance of SuperStructure, we analysed artificial datasets consisting of inter-connected clusters of localisations on a 2D plane (see Fig.2A). Clusters are homogeneously and randomly positioned on the plane with a cluster density $\rho_{cl} = 8.2\,\mu m^{-2}$ that is comparable to that of some nuclear proteins (see below). Every cluster has average radius $R_{cl} \sim 40\,nm$ and an overall internal localisation density $\rho_{em} = N_{em}/\pi R_{cl}^2 = 16000\,\mu m^{-2}$, where N_{em} is the number of

localisations per cluster. Pairs of clusters are connected with probability p_r by a sparse points distribution and only if the distance between the clusters is less than $b = 1\ \mu m$. These choices allow us to readily tune the degree of "connectivity" in the system by varying a single parameter p_r. A second parameter $p_{r_{conn}}$ is introduced to control the density of localisations within the connections ρ_{conn} (see Methods for details).

The length-scales associated to density of emitters inside clusters ρ_{em} and to the connections ρ_{conn} define the boundaries between the three regimes of $N_c(\varepsilon)$ (Fig.2B): (i) for $\varepsilon \lesssim 12\ nm$ the intra-cluster regime follows a Poissonian decay (Eq.1) with density parameter $\rho_{em} = 16000 \mu m^{-2}$ (as expected since it was set by construction); (ii) for intermediate values of ε the exponential super-cluster regime dominates (Eq.4) and the fusion of connected clusters takes place (see inset of Fig.2B); (iii) for $\varepsilon \gtrsim 60\ nm$ we expect to observe the coalescence of super- and non-connected clusters in a second super-cluster regime; this is captured by a second exponential for $p_r \neq 0$ (Eq.4). Conversely, for $p_r = 0$, we observe a single super-cluster regime that is well fitted by a Poissonian function with lower density (Eq.3), as it corresponds to the density of clusters rather than emitters within clusters (see dark-green curve in Fig.2B).

Examination of Fig.2B (inset) highlights the exponential behaviour of the super-cluster regime (ii) for different values of connectivity p_r. Importantly, a larger p_r results in an effectively shorter decay length -- or larger spatial rate of merging -- for the regime in which clusters merge into super-clusters. This strongly suggests that the effective decay length (or rate) mirrors the connectedness of the underlying super-structures (Fig.2C). In fact, these simulations reveal that the decay length represents the combined contribution of clusters density ρ_{cl} and connectivity p_r. A larger density of clusters can impact the decay length as much as a larger connectivity, as shown by simulations at fixed p_r and different ρ_{cl} (Figs.2D, S2A and S2B). In particular, we find that the functional form of the decay length is $\lambda \sim \rho_{cl}^{-1/2}\ p_r^{-0.3}$ (Figs.2D and E). The cluster density contribution is $\sim \rho_{cl}^{-1/2}$ as it depends on the typical distance between clusters and is relevant when comparing datasets with different cluster density. By combining SuperStructure with a cluster analysis, one can estimate ρ_{cl} and normalise λ to obtain the pure connectivity contribution in the decay length: $\lambda^* = \lambda / \rho_{cl}^{-1/2}$.

Finally, in order to characterise the contribution to the $N_c(\varepsilon)$ curves coming from the density of localisations within the connections, we further simulated SMLM datasets with a fixed, large connectivity p_r and varied the density of points in the connections by tuning $p_{r_{conn}}$ (see simulated datasets in Fig.2A and Fig.S2F). As expected, we observe a single super-cluster regime and the denser the connections the shorter the decay length. This indicates that our algorithm is not only able to describe how well clusters are connected, i.e. the number of connections per cluster, but also how strongly they are connected, i.e. how dense the connections are. These features are likely to be highly relevant for nuclear proteins.

Before applying this methodology to experimental data, we also tested the effect of random noise in the system, i.e. unconnected isolated localisations from biological or technical sources. We observed that in presence of random noise the decay of SuperStructure curves becomes Poissonian for large ε (see Fig.S2C) with an effective density ρ larger than the cluster density (see Fig.S2D). Decay lengths in the first super-cluster regime (yellow regime) are still distinguishable even in presence of noise at reasonable density (albeit smaller than the connection density), but their absolute values are altered with weakly connected systems more severely affected (see Fig.S2E). These observations suggest that, as in most analysis algorithms, large noise might obscure exponential decays of connected systems. In case a single Poissonian

behaviour, or a combination of exponential and Poissonian decay, are found in the SMLM dataset, it is therefore important to combine SuperStructure with an independent cluster analysis at different lengthscales (for instance at 3 or 4 selected values of ε) and a direct observation of the dataset, in order to exclude the presence of hidden connectivity.

Quantification of Super-Structures in Nuclear Proteins

We now examine biological data and apply SuperStructure to dSTORM data acquired for three different nuclear proteins (Fig.3A and B): the serine/arginine-rich splicing factor SC35, the heterogeneous nuclear RiboNuclear Protein hnRNP-C and hnRNP-U (also known as Scaffold Attachment Factor A, SAF-A). These proteins are abundantly expressed in the nucleus of human cells and are involved with RNA processing at different stages. SC35 is necessary for RNA splicing while hnRNPs are implicated in regulation and maturation of mRNA but also in chromatin structure (Nozawa et al., 2017; Xiao et al., 2012; Caudron-Herger et al., 2011). In particular, SAF-A is thought to form a dynamic homogeneous mesh that regulates large-scale chromatin organisation by keeping gene-rich loci in a decompacted state (Nozawa et al., 2017; Michieletto and Gilbert, 2019). Hence, capturing the organisation of this protein beyond the traditional single-cluster analysis is an important step towards understanding how it regulates chromatin structure in different cell stages and conditions.

Curves obtained from SuperStructure analysis after masking signal in the nuclear region are shown in Fig.3C, where we highlighted the super-cluster regimes discussed above. Global nuclear analysis is represented by filled curves, while analysis on localised ROIs by dashed ones (hnRNP-C nuclear mesh and SC35 speckles). Both hnRNPs display a first super-cluster regime for which the curves decay as exponentials, suggesting that within this range distinct clusters are in reality connected. Interestingly, while SAF-A displays a unique long super-cluster regime, hnRNP-C seems to also show a second exponential regime (filled curve). However, this regime appears at very large values of ε and is due to sparse clusters of localisations in the nucleolus. Running SuperStructure on ROIs masking out the nucleolus (dashed line) indeed generates a single exponential function, confirming that hnRNP-C clusters are fully connected. We can therefore conclude that both hnRNPs exhibit a single exponential regime, typical of fully connected meshes. On the other hand, SC35 displays exponentials with different characteristic decay rates in two distinct and significant super-cluster regimes (filled curve): one for intermediate $\varepsilon \in [10,20]\, nm$, when clusters inside speckles merge (first super-cluster regime), and another one for large $\varepsilon \in [40,150]\, nm$ indicating that speckles merge together and with isolated clusters (second super-cluster regime). The SC35 connectivity is further confirmed by running SuperStructure on ROIs masking the speckles, as we observed a clear single exponential decay (dashed line). These regimes are further confirmed by directly looking at the arrangement of identified clusters for certain values of ε (see Fig.3A inset and 3B).

From the SuperStructure curves, we first obtained the density of intra-cluster emitters by fitting the intra-cluster regime with the Poisson function (Eq.1). Interestingly, both SAF-A and SC-35 form clusters with similar densities, while hnRNP-C clusters are less dense (see Fig.3D and E). Then, in order to have a quantitative description of the clusters/speckles connectivities, we fitted the curves in the

exponential regimes (Eq.4) to extract the decay length λ. However, a direct comparison is possible only by normalising decay lengths by the cluster/speckle density (see Methods for details and Fig.S3A and B). Fig.3F highlights that while hnRNP-C has a short normalised decay length λ^* due to the highly connected clusters, SAF-A displays a weaker decay (larger λ^*) due to sparser connections. Finally, SC35 displays one (intra-speckle) very connected, even more than that of hnRNPs (small λ^*) followed by a regime (inter-speckle) that is much slower and so more weakly connected than that of hnRNPs.

In summary, our analysis revealed that while different nuclear proteins may have similar cluster sizes or densities of emitters within clusters (e.g., SAF-A and SC35) they have distinct super-cluster arrangements and connectivities. For instance, we find that the super-structures inside nuclear speckles are more connected than those formed by hnRNPs and also denser (see Figs.3E, 3F and Table SI). We stress that these features, which we further verified not emerging from technical artefacts (see Fig.S3C), cannot be quantified using standard clustering algorithms or pair-correlation functions. Additionally, the analysis in Fig.3E and F shows that our method is sensitive enough to distinguish connectivity features of two closely related wild-type hnRNPs in cell-based experiments.

The results presented in Fig.3 not only give us confidence that SuperStructure can be applied to a variety of nuclear wild type or mutated proteins in different cells, cell stages and conditions but that it also has the capability to extract unique features that may yield new mechanistic insights into the functioning of such proteins. For instance, the analysis of SC35 reveal that speckles are themselves made of clusters that are as heavily inter-connected as the clusters formed by hnRNP proteins. Given the fact that all these proteins interact with RNA, our findings suggest that RNA-binding may facilitate the formation of connections between clusters of proteins; in turn, this also points to a suspected structural role of non-coding RNAs in structuring the organisation of the nuclear interior (Hall and Lawrence, 2016). Studying the effect of RNA depletion on the super-cluster connectivity is therefore a natural next step to perform in the future.

In general, while certain mutations or conditions may not alter the size of protein cluster itself, they may affect the connectivity between clusters. In these cases, the analysis provided by SuperStructure would be invaluable and indeed essential to reveal the underlying mechanisms that guide the formation of such protein assemblies.

Ceramides clusters at the plasma membrane are not connected

To test our algorithm on a different class of molecules, we applied SuperStructure on published dSTORM datasets (Burgert et al., 2017) taken on ceramides -- membrane lipids involved in cellular trafficking (Fig.4A). The authors (Burgert et al., 2017) found that bSMase treatment increases the size of ceramides clusters and the overall localisation density. By applying SuperStructure analysis (Fig.4B), we confirmed these results and further detected that the difference in localisation density persists inside clusters (see Figs.4C, 4D, S4C and S4D). Furthermore, we detected the absence of connectivity between clusters, as the large ε regime is well-captured by a Poisson function (Eq.3), and not by an exponential (see Fig.4B and E). In other words, clusters of ceramides behave as unconnected, uniformly and randomly distributed emitters. The possibility of local connectivities at intermediate ε has been also ruled out as no

merging of clusters was observed (see Fig.S4A and B). The crossing of the curves at $\varepsilon \simeq 25\ nm$ is a consequence of the overall difference in localisation density (which in turn causes a horizontal shift between the curves, see Fig.4B inset and 4C), rather than a difference in local connectivities. The notable absence of connections between clusters of ceramides further supports that the ones detected in hnRNP-U/C and SC35 are significant.

Limitations and potential interpretation pitfalls

While we have provided evidence that SuperStructure can detect connected clusters and distinguish them from noise (at low density) or unconnected but dense clusters, in this section we discuss potential pitfalls and interpretation issues.

First, as mentioned earlier, datasets should always be segmented in order to identify the main region of interest (ROI). Spurious localisations outside the ROI (for instance outside of the nucleus, if we are interested in nuclear proteins) may affect the curves generated by SuperStructure and render their interpretation difficult. An analogous issue may arise if the localisations are embedded within heterogeneous structures, as in the case of SC35 proteins which form strongly connected structures within nuclear speckles and weakly connected outside (see Fig.3). Due to this mixed behavior over similar length-scales it is recommended to restrict the analysis to regions that display similar structural phenotypes. Even better, and to be preferred when possible, is to label the region or structure of interest with orthogonal markers.

The key difference between connected and unconnected (albeit possibly more clustered) structures is the functional form of the SuperStructure curves. However, in some cases Poisson curves may be difficult to distinguish from exponentials (especially over short intervals). In this case the best way to identify connected clusters (and distinguish them from noisier or more clustered sub-regions) is to restrict the analysis over smaller ROIs to clear potential contaminations and to additionally perform goodness-of-fit tests on the curves. Additionally, in these complex cases we also suggest to perform an independent cluster analysis over different length-scales and to directly observe datasets distributions.

As with all computational algorithms, the danger of incorrect interpretation can be addressed with quality control. In the case of SuperStructure this means directly monitoring the formation of connected clusters/structures while increasing ε. Nonetheless, thanks to its parameter-free execution, SuperStructure may offer one of the safest ways to currently analyse SMLM data.

Discussion

In this work we have introduced a novel algorithm that extends the traditional idea of cluster analysis of SMLM data and that can quantify both the connections between clusters and the density of emitters within clusters. SuperStructure introduces for the first time the concept of “connectivity” between clusters, which is different from a random distribution of points at high density. In this concept, connection points are preferentially found in between clusters and this feature manifests itself in SuperStructure curves behaving as single exponentials rather than Poissonian. Because SuperStructure is parameter-free, it does not require any prior knowledge of

the sample and it thus takes a crucial step towards a more standardised, portable and democratic quantification of complex patterns and super-structures in SMLM data.

Here, we have tested the capabilities of SuperStructure first on simulated datasets, where we observed that it could capture not only the degree of connectivity between clusters, but also the strength of the connections, and then on biological dSTORM data from nuclear proteins and membrane lipids. SuperStructure allowed us to discover that the speckles formed by the splicing factor SC35 are made of connected clusters. Further, that the density of emitters in those clusters is high and the connectivity between clusters even higher than that of hnRNP proteins. We argue that this may reflect the RNA-binding feature that characterises both hnRNPs and SC35 and that may be driving the formation of inter-connected nuclear super-structures. We highlight that this discovery could not be made simply by looking at clustering with traditional algorithms, as both proteins display clusters of similar size at small/intermediate ε.

We further stress that SuperStructure is perfectly suited to compare different datasets without a priori assumptions (albeit, as discussed before, segmentation to ROIs is recommended for strongly heterogeneous structures). The datasets of nuclear proteins we chose to analyse are an example of this. SAF-A, hnRNP-C and SC-35 are three nuclear proteins involved in the metabolism of RNA at different stages and they display three different connectivity phenotypes, which point to three different nuclear functions. In particular, SAF-A, which also plays a major role in maintaining the chromatin active loci in a decompacted state, is detected as a fully connected mesh. This finding is in agreement with a previous study that hypothesised the formation of a dynamic and RNA-interacting nuclear mesh made by SAF-A (Nozawa et al., 2017). We thus argue that SuperStructure is a useful tool for studying the structural and functional properties of this nuclear mesh. For instance, we expect that in absence of RNA, the SAF-A mesh would be disrupted and its connectivity strongly weakened (not necessarily affecting the protein clusters, which may be formed via an RNA-independent mechanism, such as phase separation by weak unspecific interactions of SAFA's intrinsically disordered domain). In turn, the application of SuperStructure would in this case be indispensable for understanding the link between the spatial arrangement, mechanics and function of this nuclear protein. A similar example is given by the V(D)J locus, whereby interacting segments appear to be trapped by a protein or chromatin network whose (super-)structure is still poorly understood (Khanna et al., 2019). We argue that SuperStructure can shed light also on this problem.

In addition to all this, super-resolved chromatin tracing (Boettiger et al., 2016; Bintu et al., 2018) and ATAC-PALM (Xie et al., 2020) generate complex datasets that will benefit from "beyond-traditional-clustering" algorithms. Connections between nano-domains and chromatin paths, do not resemble the structure of isolated clusters, but rather that of a mesh of clusters, which would be perfectly suited for quantification via the SuperStructure algorithm.

The use of SuperStructure is not limited to biological applications, and we propose it can be used as a standardised and parameter-free tool for assessing imaging technical aspects (van de Linde and Sauer, 2014; Hennig et al., 2015). One of the main issues in SMLM data, especially in dSTORM, is the evaluation of fluorophore blinking quality, as it strongly affects the localisation accuracy in the analysis process. For example, an elevated blinking frequency would result in a high emitters density (per frame) and therefore in a high localisation inaccuracy due to overlapping emissions. A similar detrimental effect could also be due to a poor blinking

signal (few emitted photons per blinking event). As a consequence, lower localisation precision of emitters may create pseudo-clusters, as well as pseudo-connections. We envisage that SuperStructure would be well suited to evaluate the blinking quality of fluorophores, for instance by measuring the emerging pseudo-connectivity in a controlled setup, such as fluorophores attached to a grid.

As discussed above, SuperStructure has been developed with the aim of going beyond "simple clustering" and in particular to measure connectivity between clusters. However, our method might be used in combination with other pair-wise distance and clustering methods. For instance, one can compute Ripley's (pair-wise distance) functions to preliminarily detect if localisations are uniform or clustered and, in case, what is the average cluster radius. Yet, Ripley's functions cannot identify single clusters or complex structures. Thus, one could use SuperStructure to determine whether the system under investigation displays connected or isolated clusters. At the same time, by computing SuperStructure curves, one can have a firm ground to decide the value of ε that can be used as input in DBSCAN for cluster analysis. This second approach can be used, for example, to measure the size or shape of local super-structures. Indeed, one can fix ε at the value that identifies super-structures, perform a cluster analysis and calculate the gyration tensor of the identified clusters.

We tested the segmentation capabilities of the latter approach by estimating the radius and circularity of SC35 speckles; we observed that it yields similar results as the well-known SR-Tesseler software (Levet et al., 2015) (see Fig.S5). Albeit SuperStructure lacks a Graphical User Interface, it has several advantages. Firstly, the analysis is OS-independent and can be easily automatised to run on a large number of cells. Secondly, since based on DBSCAN, the algorithm scales as $n_\varepsilon N^2$ in its simplest implementation (where n_ε is the number of ε values used in the analysis and N the total number of localisations). Yet, calculations on different ε are independent and so SuperStructure scales extremely well with the number of CPUs available. For instance, the analysis of $n_\varepsilon = 100$ values and 10^5 localisations can be done on a 6-core CPUs machine in about 19 minutes. Thirdly, since our algorithm is aimed at extracting "beyond-simple-clustering" information, it is flexible and intended to be used in combination with other pair-correlation or segmentation methods that are extensively employed for single-clustering analysis.

We conclude highlighting that SuperStructure provides an unbiased and parameter-free estimation of (i) density of localisations within single clusters and (ii) formation of super-structures made of connected clusters. Here we tested SuperStructure both on in simulated and cell-based SMLM datasets. Importantly, we revealed previously undocumented system-spanning structures made of connected clusters of nuclear proteins that we argue may have a functional role in shaping genome organisation. The use of SuperStructure on cells under different conditions or with protein mutations is thus an exciting direction to uncover the biological significance of these newly discovered nuclear structures.

Material and Methods

SuperStructure algorithm

SuperStructure is an algorithm that detects and quantifies super-structures formed by inter-connected clusters on SMLM datasets. Additionally, it can also evaluate the density of emitters inside clusters.

SuperStructure is mainly based on DBSCAN, a density-based algorithm to detect clusters of points in arbitrary dimensional space. The key concept underlying DBSCAN scheme is that it groups together points at high density, while it marks as outliers points in low density regions. After defining a neighbourhood size ε, a point x can be part of a cluster if the number of points $N(\varepsilon, x)$ within a circular region $\Omega(\varepsilon, x)$ of size ε centred in x, exceeds some threshold N_{min} (or is within the region $\Omega(\varepsilon, y)$ of another point y satisfying this condition).

The concept of clusters is subject to the choice of ε and N_{min} and therefore to some sort of likeness or proximity. Furthermore, the *change* in number of clusters detected by DBSCAN when varying ε contains some information of the underlying distribution of points that has been overlooked.

SuperStructure progressively runs DBSCAN to detect the number of clusters N_c within a broad range of the neighbourhood parameter ε, while N_{min} is kept fixed. The resulting $N_c(\varepsilon)$ curves, and in particular the change $dN_c(\varepsilon, N_{min})$ due to a small change in neighbourhood parameter $d\varepsilon$, contain fundamental information about the formation and organisation of super-structures and connected clusters.

As we aim for a parameter free algorithm, without losing generality, we fix $N_{min} = 0$, which means no minimum number of other emitters necessary in the neighbourhood to define a localisation as part of a cluster. For $\varepsilon = 0$, any point is found to be a cluster by itself. Then, points merge upon increasing $\varepsilon \rightarrow \varepsilon + d\varepsilon$, resulting in $dN_c/d\varepsilon \leq 0 \; \forall \, \varepsilon$. Additionally, the larger $|dN_c/d\varepsilon|$, the more identified clusters are coalescing together for a certain ε.

At ε smaller than the typical (true, rather than the one detected by DBSCAN) cluster size, the decay of $dN_c/d\varepsilon$ is determined by the intra-cluster density of points ρ_{em} (intra-cluster regime), as they are the points at the highest density. The decay of this regime is gaussian and it is described by the Poisson Function:

$$N_c(\varepsilon) = \sum_{k=0}^{m} c_k \frac{(\pi \rho_{em} \varepsilon^2)^k}{k!} \, e^{-\pi \rho_{em} \varepsilon^2} \qquad (1)$$

In order to understand the origin of this functional form, let's imagine to apply SuperStructure algorithm by setting $N_{min} = 0$ and by increasing the radius ε. For sufficiently small ε, every point is considered as a single cluster itself, as no other points are detected in its neighbourhood. However, by increasing ε, the probability of finding another point in the neighbourhood increases, implying that points start to merge in bigger clusters for small ε. It is then legitimate to argue that the number of detected clusters N_c decreases (with ε) as the probability of not finding any other emitter in the neighbourhood. This is the so-called Poisson Avoidance Function $N_c(\varepsilon) = P(n(\varepsilon) = 0) = e^{-\pi \rho_{em} \varepsilon^2}$ and it is a good approximation for very small ε, where the contribution of clusters formed by 2 emitters dominates over clusters formed by 3

or more points. For larger ε, this function underestimates the number of detected clusters. The number of detected clusters can therefore be described by the probability of not finding more than m particles in the circle of radius ε. The function we are seeking is the linear combination of the probabilities of not finding any other point in the neighbourhood and finding one or more other points (up to $m-1$). Being the probability of finding k particles $P(n(\varepsilon)=k)=\frac{(\pi\,\rho_{em}\,\varepsilon^2)^k}{k!}e^{-\pi\rho_{em}\varepsilon^2}$, it is then straightforward to get the functional form of Eq.1.

Note that $c_k=1/(k+1)$ in Eq.1 is to avoid overcounting clusters. In fact, if we consider two points within distance ε from each other (and hence in the same cluster), both points will count towards $P(n(\varepsilon)=1)$ so this contribution must be divided by 2, etc. Importantly, Eq.1 displays a natural length-scale $\kappa_0=(\pi\,\rho_{em})^{-1/2}$ that is intrinsically determined by the internal density of emitters ρ_{em}.Therefore, ρ_{em} is a parameter that can be quantified by fitting the $N_c(\varepsilon)$ curve and it can also be used to quantify the approximate upper limit of this regime (with 99% confidence level):

$$\varepsilon^* \simeq 3\kappa_0 = 3/\sqrt{\pi\,\rho_{em}} = 3R_{cl}/\sqrt{N_{em}} \qquad (2)$$

where R_{cl} is the average cluster radius and N_{em} is the average number of localisations within a single cluster. We successfully tested that SuperStructure curves are well-fitted by Eq.1 up to $m=2$ using a system where we simulated localisation of points inside a single cluster (see Fig.S1).

At ε of the order than the typical (true) cluster size, the decay is determined by the rate at which distinct clusters merge upon $\varepsilon \to \varepsilon + d\varepsilon$ (first super-cluster regime). This merging can be either due to (i) distinct clusters starting to overlap as their distance is smaller than ε or (ii) the presence of points -- which we call “connections” -- bridging two clusters. In case of total absence of connectivity and a homogeneous clusters distribution, the merging is only due to the random positioning of clusters and therefore it also follows a Poisson Function:

$$N_c(\varepsilon) = f\sum_{k=0}^{m} c_k\frac{(\pi\,\rho_{cl}\varepsilon^2)^k}{k!}e^{-\pi\rho_{cl}\varepsilon^2} \qquad (3)$$

where f is a normalisation factor and ρ_{cl} the density of clusters. We observed that SuperStructure curves of simulated systems are well-fitted by using $m=1$. This equation holds also in presence of noise, but in that case $\rho_{cl} \to \rho_{cl}+\rho_{noise}$ (see Fig.S2). The decay is different in presence of connections between clusters: connected clusters will merge at smaller ε than unconnected ones (assuming same distance between the centres of clusters). In particular, the larger the number of connections or of the local density of connection points ρ_{conn} (i.e. thicker connections), the faster the merging of bridged clusters as a function of ε and thus the larger $|dN_c/d\varepsilon|$. The functional form of this second regime is exponential in presence of connections:

$$N_c(\varepsilon) = g\cdot e^{-\varepsilon/\lambda} \qquad (4)$$

where g is a normalisation factor and λ the decay length quantifying the rate of decay, and therefore the connectivity. This decay length can be used to discern systems that exhibit either different grades of connectivity or homogeneous meshes at different

densities. Note λ purely quantifies the connectivity only when the cluster density ρ_{cl} is small and homogeneous, as we could have underlying highly dense clusters overlapping and therefore merging. We showed that $\lambda \sim \rho_{cl}^{-1/2}$ and therefore the pure connectivity decay length can be further evaluated if the density of clusters is known: $\lambda^* \sim \lambda/\rho_{cl}^{-1/2}$.

We need to stress that by choosing $N_{min} = 0$ connections will also be considered as points to be merged. However, it is important that we identify "connection" points as having a lower local density ρ_{conn} than the groups of points that are bridged by them (clusters). In this way, they will merge in this second regime to form super-structures. The limiting case in which the local density of connection points is the same as the one in the clusters at the two ends of the connections is indistinguishable from the case of one elongated cluster. A special case is that in which both clusters and connections have the same density of points, but the connections are slightly detached from the clusters, thus forming three independent clusters at intermediate ε which may then merge (we assume this to be a rare event). The above reasoning can be extended to multiply connected clusters via the analysis of pair-wise connections.

At larger ε, we could have additional super-clusters regimes if the system is heterogeneous. Most common cases showing two (or more) super-cluster regimes are the following: (1) inhomogeneous system displaying different connectivities at different lengthscales, (2) connected clusters embedded in a noisy environment (in this case we observe an exponential followed by a poissonian decay) and (3) unconnected clusters within a random noise and/or unconnected clusters at different densities (in this case we observe two or more poissonian decays).

SuperStructure Pipeline

In order to apply SuperStructure, we adopt the following steps:

1. Generation of SuperStructure curves. We run SuperStructure on a SMLM dataset by first masking our data in the region of interest (ROI), such as the nucleus for nuclear proteins as mentioned in the section below. Then, we choose a ε-range to analyse. For example, in SMLM datasets of nuclear proteins a typical choice is $\varepsilon \in [0{:}\,200]\ nm$ with $d\varepsilon = 2\ nm$. One should notice that lower $d\varepsilon$ may be necessary for fitting the intra-cluster regime. SuperStructure curves are generated by progressively running DBSCAN clustering algorithm on the SMLM dataset in the chosen ε-range (and $N_{min} = 0$). The DBSCAN software we use is from https://github.com/gyaikhom/dbscan and the progressive run is performed with bash scripts available in the repository. SuperStructure output curves are saved in a three-columns file $(\varepsilon, N_{cl}, N_{cl}/N_{loc})$, where N_{cl} is the number of detected clusters for the corresponding ε and N_{loc} the number of total localisations. Additionally, the classification of localisations in clusters is saved on a separate file for every ε.

2. Evaluation of SuperStructure regimes. As a second step, we evaluate regimes by plotting and investigating SuperStructure curves (we adopt a log-scale in the y-axis). This step includes a preliminary check for the number of regimes and their decay behaviour (exponential vs. poissonian). In the case we observe a single Poissonian behaviour, we can state that the dataset does not show any, or very limited, connectivity, and therefore we are in presence of homogeneous isolated clusters (and

eventually noise). Limited connectivity needs to be checked with a cluster analysis and direct dataset observation in case noise has obscured an exponential decay. On the other hand, if we observe a single exponential regime (a straight line in a log-linear plot) we conclude that the system is made of fully connected clusters. If SuperStructure curves show multiple super-cluster regimes, it is likely that the system is heterogeneous. Indeed, multiple exponential regimes may reflect heterogeneous/multi-scale connectivities combined with heterogeneous distributions of clusters. Alternatively, we may find also a combination of exponential and poissonian regimes and in this case the system may be made of connected clusters embedded in a noisy region. Other more complex combinations may be possible; however, one should notice that in heterogeneous systems it might be difficult to recognise and fit super-cluster regimes. To clarify these contributions, it is useful to combine the analysis of SuperStructure curves with a direct observation of the dataset and identified structures and to run SuperStructure on smaller ROIs to analyse different regions of the sample with similar structural phenotypes. Nonetheless, SuperStructure will be able to unambiguously detect differences in connectivity and behaviours in, e.g., samples that have been subjected to different conditions or expressing mutated proteins.

3. Fit of SuperStructure regimes. Once regimes have been identified, one needs to define the boundaries where regimes crossover from one to another. This can be either done manually or by using a pre-emptive goodness-of-fit test (this procedure would also define fitting ranges). The intra-cluster regime is typically fitted with a Poisson Equation (Eq.1) to evaluate the density of emitters inside clusters as well as to obtain an estimation of the upper limit of the intra-cluster regime (using Eq.2). For super-cluster regimes, we use Eq.3 if they show a Poissonian decay (curved on a log-linear plot) or Eq.4 if they otherwise appear straight on a log-linear plot; from the latter, we quantify the connectivity parameter λ. We can then additionally calculate the cluster density ρ_{cl} to extract the pure connectivity part $\lambda^* = \lambda/\rho^{-1/2}$. The cluster density ρ_{cl} can be computed by performing a cluster analysis with DBSCAN on local circular regions representative of that decay regime and by fixing ε at the start of that regime. For instance, by counting the number of clusters one obtains by fixing ε at the beginning of the yellow area in Fig.3. In the section below and in Fig.S3, we describe in detail the procedure for λ normalisation for the nuclear proteins' datasets. Finally, and optionally, it is also possible to define a single function fitting the entire curve by either (a) defining a piecewise function where every "piece" is the fit of the corresponding regime or (b) adding together the contribution of the different regimes (appropriately weighted). We performed fits with a combination of bash and gnuplot scripts available in the repository.

Simulated datasets generation and SuperStructure analysis

The simulated dataset consists of spatially homogeneous and interconnected clusters randomly distributed on a plane. We set to work with clusters made by taking random clusters centres on the plane and by sampling $N_{em} = 80$ emitters within a Gaussian of standard deviation $\sigma_{em} = 20\,nm$, thereby setting the cluster radius to $R_{cl} = 2\,\sigma_{em} = 40\,nm$ with a 95% confidence and the intra-clusters emitters density at $\rho_{em} = 16000\,\mu m^{-2}$. The clusters are positioned in a $L = 3.5\,\mu m$ large area and their number N_{cl} is varied in order to consider different clusters densities. In the example shown in

the main text, we fixed $N_{cl} = 100$ thus fixing a cluster density to about $\rho_{cl} = 8.2\ \mu m^{-2}$ roughly similar to the values found in experiments for some nuclear proteins. Pairs of clusters are connected with probability p_r if they are positioned closer than a distance $b = 1\ \mu m$. The value of p_r is calculated as the ratio between the actual drawn connections and $N_{cl}(N_{cl} - 1)/2$, which is the maximum possible connections (i.e. when every cluster is connected with every other cluster). In order to generate a single connection, we considered the vector joining the centres of two clusters and sampled one emitter with probability $p_{r_{conn}}$ every $10\ nm$. Emitters are sampled from a 2D gaussian centred on the vector connecting the two clusters centres and with a width $\sigma_{conn} = 10\ nm$. In the main text we fixed $p_{r_{conn}} = 0.5$. Note that p_r controls the number of connections, while $p_{r_{conn}}$ their density ρ_{conn}. We generated at least 20 independent replicas for each simulated dataset using a combination of bash and python scripts, then we run SuperStructure analysis in the range $\varepsilon \in [0{:}400]\ nm$ with a change $d\varepsilon = 2\ nm$. If not differently specified, the first super-cluster regime was fitted with Eq.4 for $\varepsilon \in [15{:}60]$, while the second super-cluster regime either with Eq.3 (unconnected systems) or Eq.4 (connected systems) for $\varepsilon \in [70{:}300]$.

Experimental details for generating experimental dSTORM dataset for SAF-A, hnRNP-C and SC-35

Cells Preparation for dSTORM imaging. hTERT-RPE1 cells (ATCC, cat# ATCC-CRL-4000) were grown overnight in an 8-well Lab-Tek II Chambered Coverglass -- 1.5 borosilicate glass (Thermofisher scientific) at 37 degrees at initial concentration of $10^5\ cells/ml$ in $400\ \mu l$ ($\sim 40\%$ confluency). We fixed the cells with 4% PFA (Sigma-Aldrich) for 10 minutes, followed by wash in PBS, permeabilisation with 0.2% Triton X-100 (Sigma-Aldrich) for 10 minutes, washed in PBS again and blocked with 1% BSA (Sigma-Aldrich) for 10 minutes.

Immuno-fluorescence labelling was done by exposing the cells for 2 hours to (i) hnRNP-U polyclonal rabbit antibody (A300-690A, Bethyl Laboratories) at $10\ \mu g/ml$ or (ii) hnRNP-C1/C2 (4F4) mouse monoclonal antibody (sc-32308, Santa Cruz Biotechnology) at $0.2\ \mu g/ml$ or (iii) SC-35 mouse monoclonal antibody (ab11826, abcam) at $2\ \mu g/ml$ and then washed. Then, cells were exposed for 1 hour to secondary antibody. The secondary antibody was made by AffiniPure $F(ab')_2$ Fragment Donkey Anti-Rabbit or Donkey Anti-Mouse IgG (H+L) (711-006-152 and 715-007-003, Jackson ImmunoResearch Europe Ltd) conjugated to the organic fluorophore CF647 (92238A-IVL, Sigma-Aldrich) at a stechiometric ratio of about 1.

Oxygen scavenger imaging buffer for dSTORM was prepared fresh on the day and the recipe employed was similar to that of (McSwiggen et al., 2019). We mixed (i) $5.3\ ml$ of $200\ mM$ Tris and $50\ mM$ NaCl solution with (ii) $2\ ml$ of 40% glucose solution, (iii) $200\ \mu l$ of GLOX, (iv) $1.32\ ml$ of $1M$ 2-mercaptoethanol (Sigma-Aldrich) and (v) $100\ \mu l$ of $50\ \mu g/ml$ DAPI solution (Sigma-Aldrich). The GLOX solution was made by mixing $160\ \mu l$ of $200\ mM$ Tris and $50mM$ NaCl with $40\ \mu l$ of catalase from bovine liver (Sigma-Aldrich) and $18\ mg$ of glucose oxidase (Sigma-Aldrich).

The $8.9\ ml$ final solution was enough to fill the chambers of the 8-well dish; a coverglass was sealed at the top of the dish to prevent inflow of oxygen.

dSTORM Acquisition. We performed 3D-STORM acquisitions using a Nikon N-STORM system with Eclipse Ti-E inverted microscope with laser TIRFilluminator (Nikon UK Ltd, Kingston Upon Thames, UK). We equipped the microscope with a CFI SR HP Apo TIRF 100x objective lens (N.A. 1.49) and applied a 1.5X additional optical zoom. We also used a cylindrical astigmatic lens to obtain elliptical shapes for emitters that reflect their z-position (Huang et al., 2008). Laser light was provided via a Nikon LU-NV laser bed with 405, 488, 561, $640\,nm$ laser lines. In particular, CF647 fluorophores were stochastically excited using the $640\,nm$ laser beam with an additional 405 weak pulse. Images were acquired with an Andor iXon 897 EMCCD camera (Andor technologies, Belfast UK). The Z position was stabilised during the entire acquisition by the integrated perfect focus system (PFS). Acquisition were performed at room temperature.
For every nucleus, we acquired a stack of $20000\ frames$ at $19\,ms$ exposure time by using the Nikon NIS-Element software. Acquired images have a $256\,x\,256$ pixel resolution with pixel size equal to $106\,nm$. For every condition (SAF-A, hnRNP-C, SC35) we acquired 6 nuclei, i.e. 6 independent datasets.

Raw images and post-processing analysis. The raw stack of frames was initially segmented based on a DAPI marker to carefully mask out the extra-nuclear signal. Then, frames were analysed using FIJI (Schindelin et al., 2012) and in particular the Thunderstorm plugin (Ovesný et al., 2014). Firstly, we filtered them by using Wavelet functions to separate signal from noise. The B-Spline order was set to 3 and the B-Spline scale to 2.0 as suggested in (Ovesný et al., 2014) for localisations of around $5\ pixels$ size. In order to localise the emitters centroids, we thresholded filtered images (threshold value was set 1.2 times the standard deviation of the 1st Wavelet function) and calculated the local maximum relative to the 8 nearest neighbours. Finally, we fitted the emitters signal distribution with elliptical gaussians (ellipses are necessary for z-position reconstruction) using the weighted least square method and by setting $3\ pixels$ as initial fitting radius and $1.6\ pixels$ as initial sigma.
Localised data was then post-processed using the same plugin. (i) We corrected the XY drift using a pair correlation analysis, (ii) filtered data with a position uncertainty $< 40\,nm$, (iii) restricted the z-position to the interval $[-100{:}\,100]\,nm$ and projected the data in a 2-dimensional plane, as the z-axis precision is around $100\,nm$.
Reconstructed images shown in the main text were created by using the average shifted histograms method of the same plugin with a $10X$ magnification ($10.6\,nm/pixel$).

SuperStructure analysis of nuclear protein data

SuperStructure analysis was run on the entire nuclear region by setting $N_{min} = 0$ and by increasing ε in the range $[0{:}\,200]\,nm$ and "all-nucleus" curves were generated for 6 independent nuclei. We set change rate $d\varepsilon = 0.25\,nm$ for $\varepsilon \in [0{:}\,10]\,nm$ and $d\varepsilon = 10\,nm$ for $\varepsilon \in [10{:}\,200]\,nm$. This choice was due to the higher resolution necessary to extract intra-cluster information at small ε. As shown in Fig.3, SuperStructure "all-nucleus" curves show that SAF-A has a single exponential super-cluster regime, while hnRNP-C and SC35 have two regimes. In the hnRNP-C case, the second regime is due to weakly connected and sparse clusters in nucleoli, while in SC35 to the cluster/connectivity heterogeneity in the system (i.e. speckles). Therefore, we

additionally run SuperStructure analysis on local regions of interest (ROIs) for hnRNP-C and SC35 to obtain the isolated contribution for the first super-cluster regime. In particular for hnRNP-C we considered 5 independent circular ROIs per nucleus with radius $r = 1.5\ \mu m$ within the nuclear mesh; for SC35, we considered 5 independent circular ROIs per nucleus with radius $r = 0.5\ \mu m$ within speckles. We run the analysis on these ROIs and generated SuperStructure "local" curves (5 for each nucleus).
The values of the intra-cluster density ρ_{em} were extracted by fitting with Eq.1 the intra-cluster regime in the "all-nucleus" curves in the range $\varepsilon \in [0{,}3]\ nm$. Resulting average values are: $\rho_{em}^{hnRNP-C} = 7973 \pm 1732\ \mu m^{-2}$, $\rho_{em}^{SAF-A} = 16998 \pm 2444\ \mu m^{-2}$ and $\rho_{em}^{SC35} = 18680 \pm 1520\ \mu m^{-2}$.
Then, we identified the super-cluster regimes of interest: the first super-cluster regimes of SAF-A and hnRNP-C, and both super-cluster regimes of SC35 (SC35-1 and SC35-2). For SAF-A and SC35-2, the decay length λ was obtained by fitting "all-nucleus" curves with Eq.4. For hnRNP-C and SC35-1 instead, we fitted the "local" curves (5 curves per nucleus) and then we averaged λ values obtained from different "local" curves in the same nucleus. Fit ranges are $\varepsilon \in [16{,}100]\ nm$ for SAF-A, $\varepsilon \in [14{,}70]\ nm$ for hnRNP-C, $\varepsilon \in [8{,}20]\ nm$ for SC35-1 and $\varepsilon \in [40{,}150]\ nm$ for SC35-2.
Finally, the values of λ for SAF-A, hnRNP-C, SC35-1 and SC35-2 were normalised by the cluster density: $\lambda^* = \lambda\ /\rho_{cl}^{-1/2}$. In the case of SAF-A and SC35-2, the normalisation was performed for λ for every nucleus by using the average cluster density ρ_{cl} of that nucleus. In particular, ρ_{cl} was calculated as the average of the cluster density in 5 independent circular regions of radius r in the same nucleus as shown in the example of Fig.S3A. In the case of hnRNP-C and SC35-1 where λ values were obtained from "local" curves, the normalisation of λ was performed using the cluster density of the same local region; then λ^* values obtained from different regions in the same nucleus were averaged (see Table SI). The number of clusters estimation (to calculate the cluster density) was made with DBSCAN by setting $N_{min} = 0$ and ε close to the beginning of the exponential regime of interest, as shown in Fig.S3B, and by keeping only clusters with at least 30 particles. In order to compute the cluster density, for SAF-A and hnRNP-C we set local circular regions of radius $r = 1.5\ \mu m$ and fixed $\varepsilon = 20\ nm$ for cluster analysis (for hnRNP-C the same local regions as defined above). For SC35, we considered two sets of local regions: (i) inside speckles to normalise the shorter decay length where we used ROIs with radius $r = 500\ nm$ and fixed $\varepsilon = 10\ nm$ for cluster analysis (same regions as above); (ii) outside speckles to normalise the longer decay length, where we used ROIs with radius $r = 1.5\ \mu m$ and $\varepsilon = 40\ nm$ for cluster analysis. Average nuclear values of λ, ρ_{cl} and λ^* are shown in Table SI.

SuperStructure analysis of ceramides data

SuperStructure analysis was run on the two ceramides datasets provided by the authors of (Burgert et al., 2017), namely +bsMase and -bsMase, by setting $N_{min} = 0$ and $\varepsilon \in [0{:}200]$. We set $d\varepsilon = 0.5\ nm$ for $\varepsilon \in [0{:}10]\ nm$ and $d\varepsilon = 2\ nm$ for $\varepsilon \in [10{:}200]\ nm$. This choice was due to the higher resolution necessary to extract intra-cluster information at small ε. From the curves in Fig.4B, it is clear that there is not any strong connectivity (we observe a Poissonian decay). Therefore, we identified free unclustered emitters as noise. We have additionally run SuperStructure in 16 independent local circular regions of radius $r = 1.5\ \mu m$ to extract the quantities of interest. In particular, we measured the average densities of total localisations: $\rho_{loc}^{+} =$

$595 \pm 130\ \mu m^{-2}$ and $\rho_{loc}^{-} = 475 \pm 87\ \mu m^{-2}$, respectively for + and - bsMase treatment. This is in accordance with results in the original paper. Then, we fitted "local" SuperStructure curves in the intra-cluster regime with Eq.1 for $\varepsilon \in [0:3]\ nm$: $\rho_{em}^{+} = 22391 \pm 3306\ \mu m^{-2}$ and $\rho_{em}^{-} = 15505 \pm 3470\ \mu m^{-2}$ repsectively for + and -bsMase treatments. Finally, we fitted "local" SuperStructure curves in the super-cluster regime with Eq.3 in the range $\varepsilon \in [50:200]\ nm$ for +bsMase and $\varepsilon \in [60:200]\ nm$ for -bsMase (the difference in fit starting value is explained by the two curves horizontal shift): $\rho_{sc}^{+} = 62.01 \pm 20.76\ \mu m^{-2}$ and $\rho_{sc}^{-} = 43.56 \pm 11.05\ \mu m^{-2}$. These values are in accordance with the sum of cluster density and noise at the ε-value were the fit starts. We have additionally performed a cluster analysis with DBSCAN and results are in agreement with the original paper results (see Fig.S4 for details). In order to verify that there is not any limited connectivity hidden by noise, we performed a cluster analysis also at two different values of ε and monitored the change in density of clusters and density of free emitters (see Fig.S4 for details).

Online Supplemental Material

Table.SI recapitulates values for λ, ρ_{cl} and λ^{*} in nuclear proteins data. **Fig.S1** shows a simulated distribution of points inside a single cluster and how it is well represented by Eq.1 in Methods. **Fig.S2** shows SuperStructure curves (or decay lengths) for simulated datasets in different conditions: **(A)** different p_r and doubling the typical cluster density; **(B)** different values of cluster density; **(C)** different p_r without and with noise addition; **(D)** unconnected clusters with noise at different densities; **(E)** decay lengths of the first super-cluster regime for different connectivities p_r as function of noise density; **(F)** homogeneous mesh (high p_r) and different values of connection density (controlled by $p_{r_{conn}}$). **Fig.S3 (A)-(B)** shows how the normalisation of λ was performed in nuclear protein data (exhaustively explained in Methods); **(C)** shows that nuclear proteins connectivity is not a technical artefact. **Fig.S4** shows that there is no local connectivity in ceramides data and confirms original paper results on ceramides cluster size. **Fig.S5** shows SuperStructure + DBSCAN segmentation capabilities by estimating the radius and circularity of SC35 speckles alongside SR-Tesseler software.

Aknowledgments

MM is a cross-disciplinary post-doctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (core grant MC_UU_00009/2 to the MRC Institute of Genetics and Molecular Medicine). SvdL is supported by the Academy of Medical Sciences/the British Heart Foundation/the Government Department of Business, Energy and Industrial Strategy/the Wellcome Trust Springboard Award (SBF003\1163). N.G. is funded by the UK Medical Research Council (MC_UU_00007/13). DM is a Royal Society University Research Fellow and was supported by the Leverhulme Trust (ECF-2019-088) and ERC StG (TAP, 947918). The authors thank the support of the Scottish University Life Science Alliance through a technology seed grant Worktribe Project ID 8824507. The authors thank the ESRIC Imaging Team (IGMM section), in particular Matthew Pearson and Ann Wheeler for their support. The authors are grateful to Markus Sauer for providing with the ceramides data. MM and DM would also like to thank Ibrahim Cissè for an igniting discussion. The authors also thank discussions with Davide Marenduzzo's group.

Data Availability

The simulated and experimental datasets that support the findings of this study are available from the corresponding authors upon request.

Code Availability

The code for the generation of SuperStructure curves is available from *https://git.ecdf.ed.ac.uk/dmichiel/superstructure*.

Author Contributions

M.M., D.M. and N.G. conceived the project. M.M. and D.M. analysed both simulated and experimental datasets. M.M., S.v.d.L. and D.M. generated the simulated dataset. M.M., E.L. and D.M. performed super-resolution experiments and localisation analysis. M.M., D.M., S.v.d.L. and N.G. wrote the manuscript with input from all the authors.

Competing Interests

The authors declare there are not competing interests.

Figures and Supplemental Figures

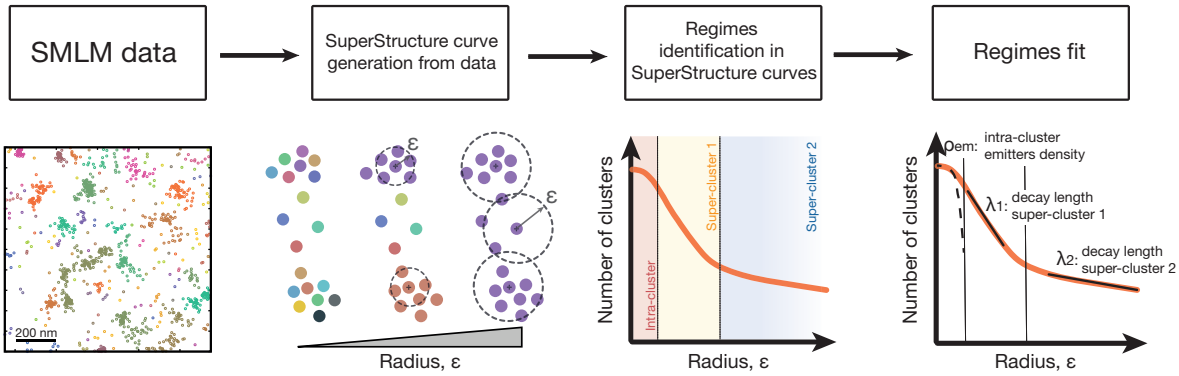

Figure 1. Working principle of SuperStructure analysis. **(1)** SMLM data is taken as input for the analysis. **(2)** Cluster analysis is run using the DBSCAN algorithm with $N_{min} = 0$ and ε progressively increasing in an adequate range for the system. SuperStructure curves describe the number of detected clusters N_c as a function of ε are generated. **(3)** SuperStructure curves are plotted and inspected to identify super-cluster regimes representing the onset of connected structures. **(4)** Intra- and super-cluster regimes are fitted with our models (see Methods) to quantify the emitters density inside clusters ρ_{em} and the connectivity among clusters (via the decay length λ_i for super-cluster regime i).

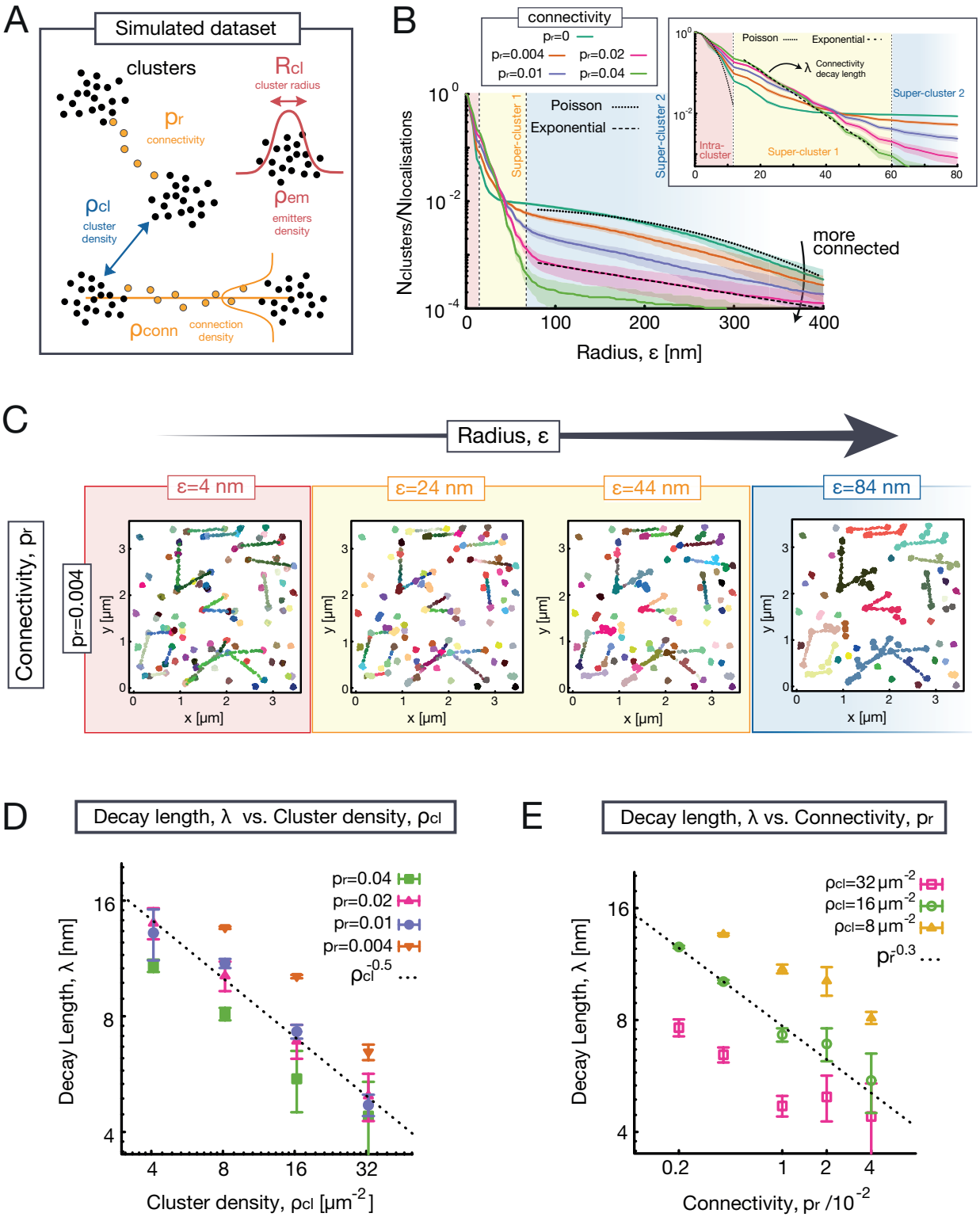

Figure 2. Evaluating SuperStructure on simulated datasets. A. Sketch representing the artificial dataset consisting of inter-connected clusters of localisations on a 2D plane. Clusters are characterised by an internal density of localisations ρ_{em} and radius R_{cl} and are randomly distributed on the plane at an average cluster density ρ_{cl}. Clusters can be connected by a sparse point distribution with probability p_r and connections have a density of points ρ_{conn} (controlled by the $p_{r_{conn}}$ parameter). **B.** Average SuperStructure curves (zoomed in the inset) for simulated datasets with different connectivity p_r. Other parameters are kept fixed: average cluster radius $R_{cl} \simeq 40\ nm$, emitters density within clusters $\rho_{em} =$

$16000\ \mu m^{-2}$, cluster density $\rho_{cl} = 8.2\ \mu m^{-2}$ and $p_{r_{conn}} = 0.5$ (which fixes the density of emitters within connections ρ_{conn}). The curves show the number of detected clusters normalised by the total number of localisations. Curves are the average of 20 independent simulated datasets. Shaded regions represent the standard deviation from the average. Three regimes can be distinguished: (i) intra-cluster (red), (ii) first super-cluster (yellow) and (iii) second super-cluster regime (blue). The decay in the intra-cluster regime corresponds to a Poisson avoidance function with density parameter $\rho_{em} = 16000\ \mu m^{-2}$ (Eq.1, dotted line in the inset). The first super-clusters regime can be fitted by a single exponential (Eq.4, dashed line in the inset) which returns an effective decay length λ. The second super-cluster regime can be fitted with another exponential if $p_r \neq 0$ (Eq.4, dashed line in the main figure). In case of $p_r = 0$, there is only one super-cluster regime and it follows a Poisson function with density parameter $\rho_{cl} = 8.2\ \mu m^{-2}$ (Eq.3, dotted line in the main figure). **C.** Snapshots of detected clusters for an artificial dataset with connectivity $p_r = 0.004$ and by progressively increasing the value of the radius $\varepsilon = 4, 24, 44, 84\ nm$. **D.** Decay length λ versus cluster density ρ_{cl} scales as $\rho_{cl}^{-0.5}$ for any value of connectivity p_r. **E.** Decay length λ versus connectivity p_r scales as $p_r^{-0.3}$ for different values of ρ_{cl}. In (**D**) and (**E**) 20 independent datasets were fitted with Eq.4 and the resulting λ values were averaged. Vertical bars represent the standard deviation from the average.

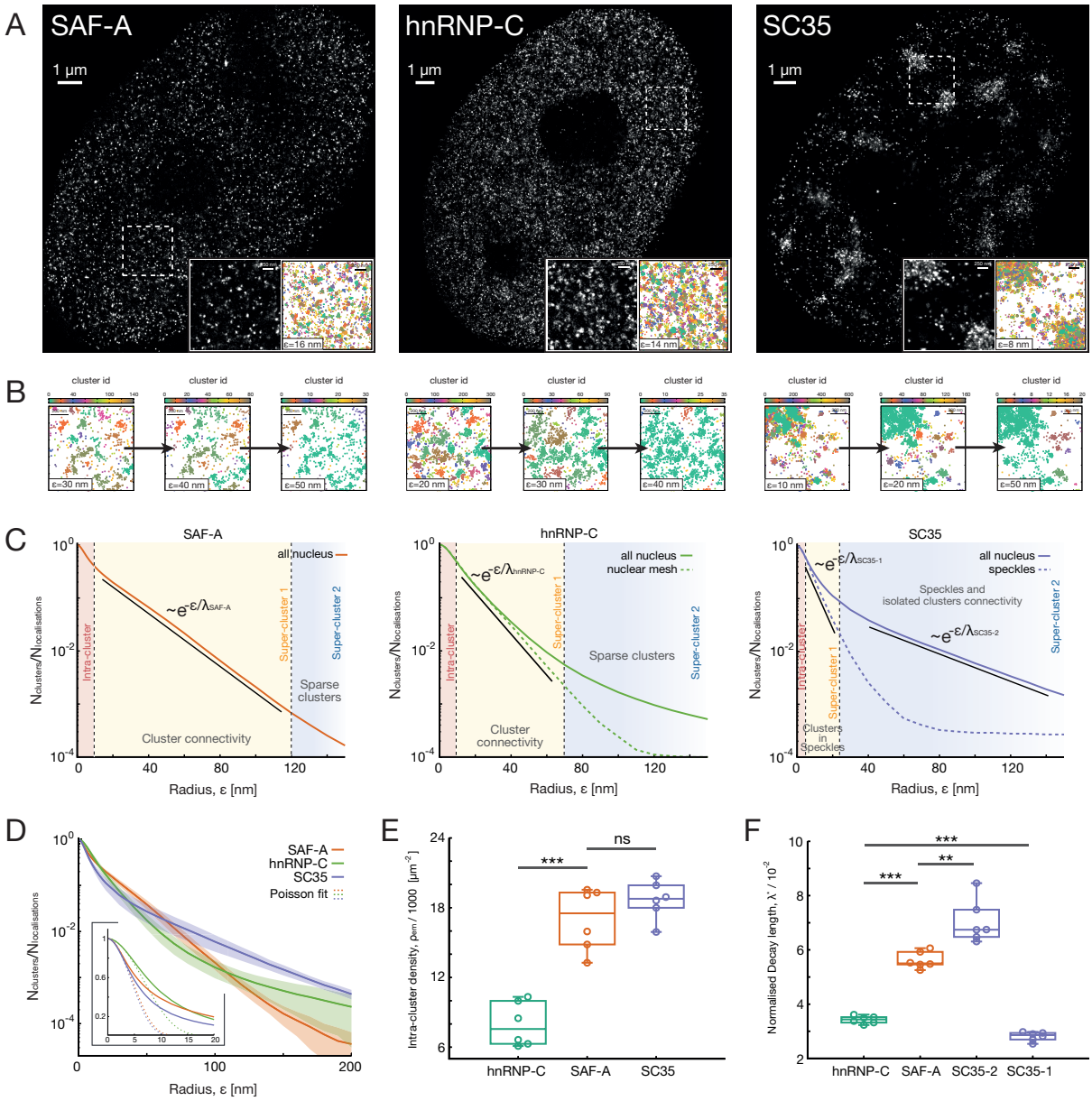

Figure 3. Application of SuperStructure algorithm to SAF-A, hnRNP-C and SC35 super-resolution data. A. Reconstructed dSTORM images by using the shifted histograms method with a pixel size of $10.6\,nm$. Insets of $4\,\mu m^2$ size of reconstructed dSTORM images and spatial positions of the data. Palettes represent the cluster id computed by running SuperStructure with $N_{min} = 0$ and ε at the start of the first super-cluster regime. **B.** Identified clusters for increasing values of ε in the regimes where clusters merge. **C.** Normalised average SuperStructure curves in the range $[0{:}150]\,nm$. The number of detected clusters has been normalised with the total number of localisations in the system. The average is calculated over 6 independent datasets (nuclei). Solid curves: SuperStructure analysis was run on the entire nucleus and the resulting curves for the 6 independent datasets were averaged ("all-nucleus" curves). Dashed curves: SuperStructure analysis was run in 5 local regions of interest (ROIs) for each of the 6 nuclei, then the curves of each region (for each nucleus) were averaged ("local" curves). In hnRNP-C these local regions were chosen within the nuclear mesh (to exclude nucleoli) and in SC35 within speckles. Vertical dashed lines highlight different SuperStructure regimes: intra-cluster, first super-cluster and second super-cluster regimes. For SAF-A and hnRNP-C the exponential regime of clusters merging (first super-cluster regime) is highlighted with a solid straight line. In case of SC35, two regimes are highlighted: the merging of clusters within speckles (first super-cluster regime) and the merging of speckles with isolated clusters (second super-cluster regime). **D.** Normalised "all-nucleus" average SuperStructure curves in the range $[0{:}200]\,nm$ for the three proteins. Average is computed

over 6 nuclei. Shaded regions represent standard deviation from the average. Poisson fits (Eq.1) for the intra-cluster regime at small ε are shown in the inset. **E.** Intra-cluster density of emitters ρ_{em} as parameter of Poisson fit for 6 independent nuclei (Eq.1). **F.** Normalised decay length λ^* for the super-cluster regimes highlighted in **C** for 6 independent nuclei. SuperStructure curves were fit with Eq.4 to extract the decay length λ, then the normalisation $\lambda^* = \lambda/\rho_{cl}^{-1/2}$ was performed (where ρ_{cl} is the detected cluster density at the beginning of each regime of interest). P-values were calculated using a Student's T-test: $ns\ P > 0.05$; $*\ P < 0.05$; $**\ P < 0.01$; $***\ P < 0.001$.

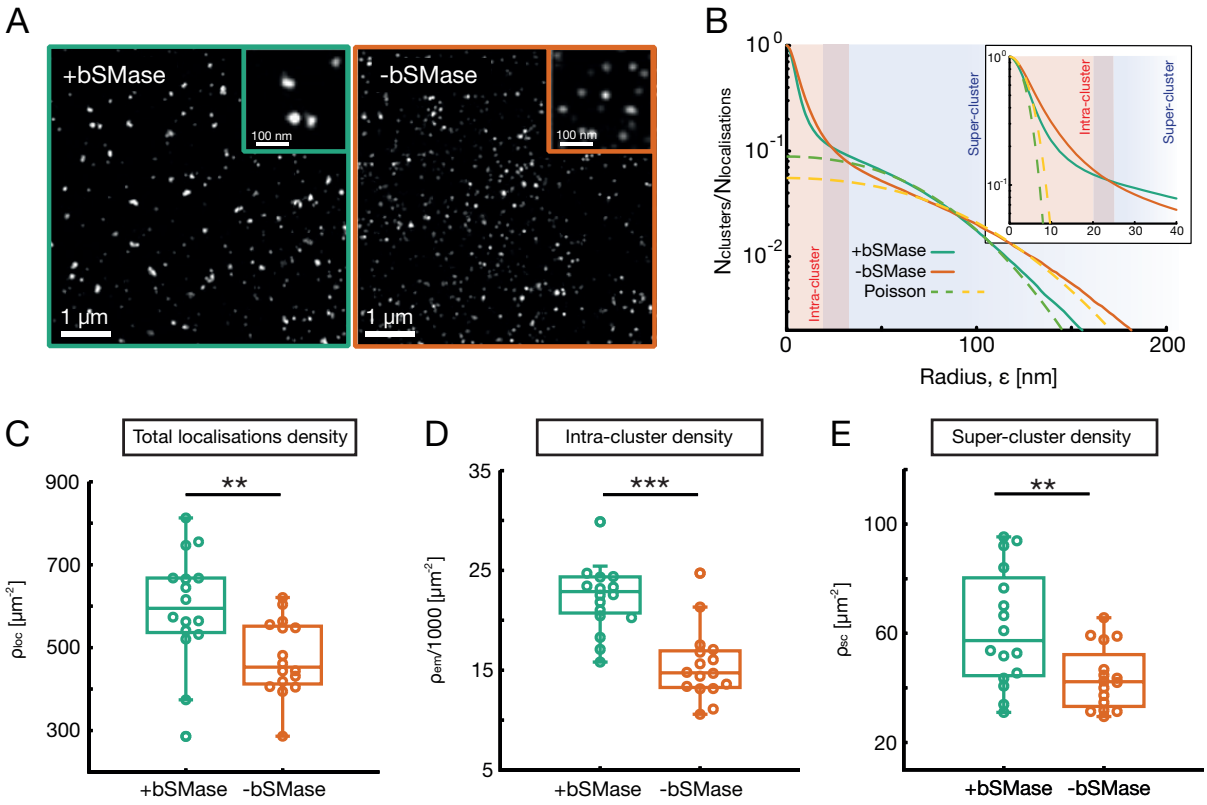

Figure 4. Application of SuperStructure algorithm to ceramides data from (Burgert et al., 2017). **A.** dSTORM reconstruction of ceramides dataset using the shifted histogram method. The left panel represents signal from cells treated with bSMase; the right panel is a control without treatment. **B.** SuperStructure curves of the two conditions for the entire dataset. Curves show the number of detected clusters normalised by the total number of localisations. The red region highlights the intra-cluster regime, while the blue region the Poissonian unconnected super-cluster regime. Shaded purple region highlights the horizontal shift between the two curves. Dashed lines represent Poisson fits at low and high ε. **C. – E.** Average density of total localisations (**C**), intra-cluster density extracted as parameter from Poisson fit (Eq.1) (**D**) and overall density in the super-cluster regime extracted as parameter from Poisson fit (Eq.3) (**E**) for + and -bsMase treatment datasets. Calculations and fits were performed on data and SuperStructure curves from 16 independent circular regions of radius $r = 1.5\,\mu m$ within the original dataset. P-values were calculated using a Student's T-test: $ns\ P > 0.05$; $*\ P < 0.05$; $**\ P < 0.01$, $***\ P < 0.001$.

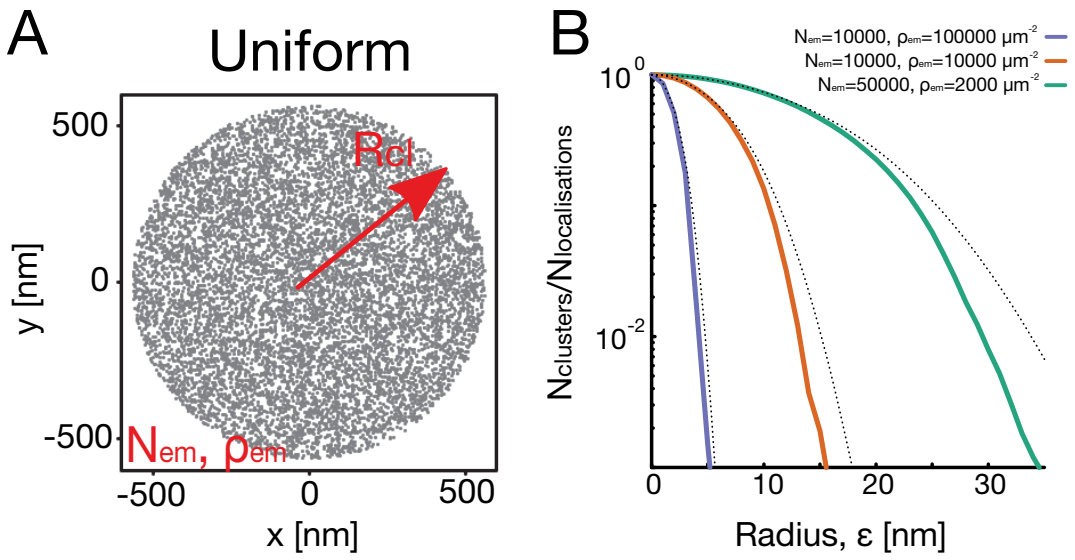

Figure S1. A. In order to test the Poissonian functional form (Eq.1) of the intra-cluster regime of SuperStructure curves, we simulated localisations inside clusters as a uniform distribution of N_{em} points distributed within a circle of radius R_{cl}. The resulting average density is ρ_{em}. The number of points included in any circular sub-region of radius ε is, on average, $n(\epsilon) = \pi\rho_{em}\epsilon^2$, and is in fact itself Poisson distributed. **B.** To check the theoretical prediction of Eq.1 we have created simulated datasets for various ρ_{em} and N_{em}. The theoretical predictions (dotted lines) with $m = 2$ are in good agreement with the SuperStructure curves, indicating that indeed Eq.1 correctly captures the behaviour of uniformly distributed points forming one idealised cluster. However, note that for $m = 2$ there is already an over-counting of clusters at large values of ϵ due to the fact that DBSCAN merges indirectly related emitters in a single big cluster. This suggests not to extend the summation to higher values of m. From Eq.1, the end of the intra-cluster regime can be approximated by the width of the Poisson function, i.e. $\varepsilon^* \simeq 3\kappa_0$ (at 99 % confidence level), where $\kappa_0 = 1/\sqrt{\pi\rho_{em}}$ is the decay length identified by Eq.1. This is confirmed by observing that predicted ε^* for the curves are $\varepsilon^*(\rho_{em} = 2000\,\mu m^{-2}) \simeq 38\,nm$, $\varepsilon^*(\rho_{em} = 10000\ \mu m^{-2}) \simeq 18\,nm$ and $\varepsilon^*(\rho_{em} = 100000\,\mu m^{-2}) \simeq 5.3\,nm$, which correspond to $N_{cl}/N_{em} \simeq 10^{-3}$ (when most of the points have been merged in a single cluster).

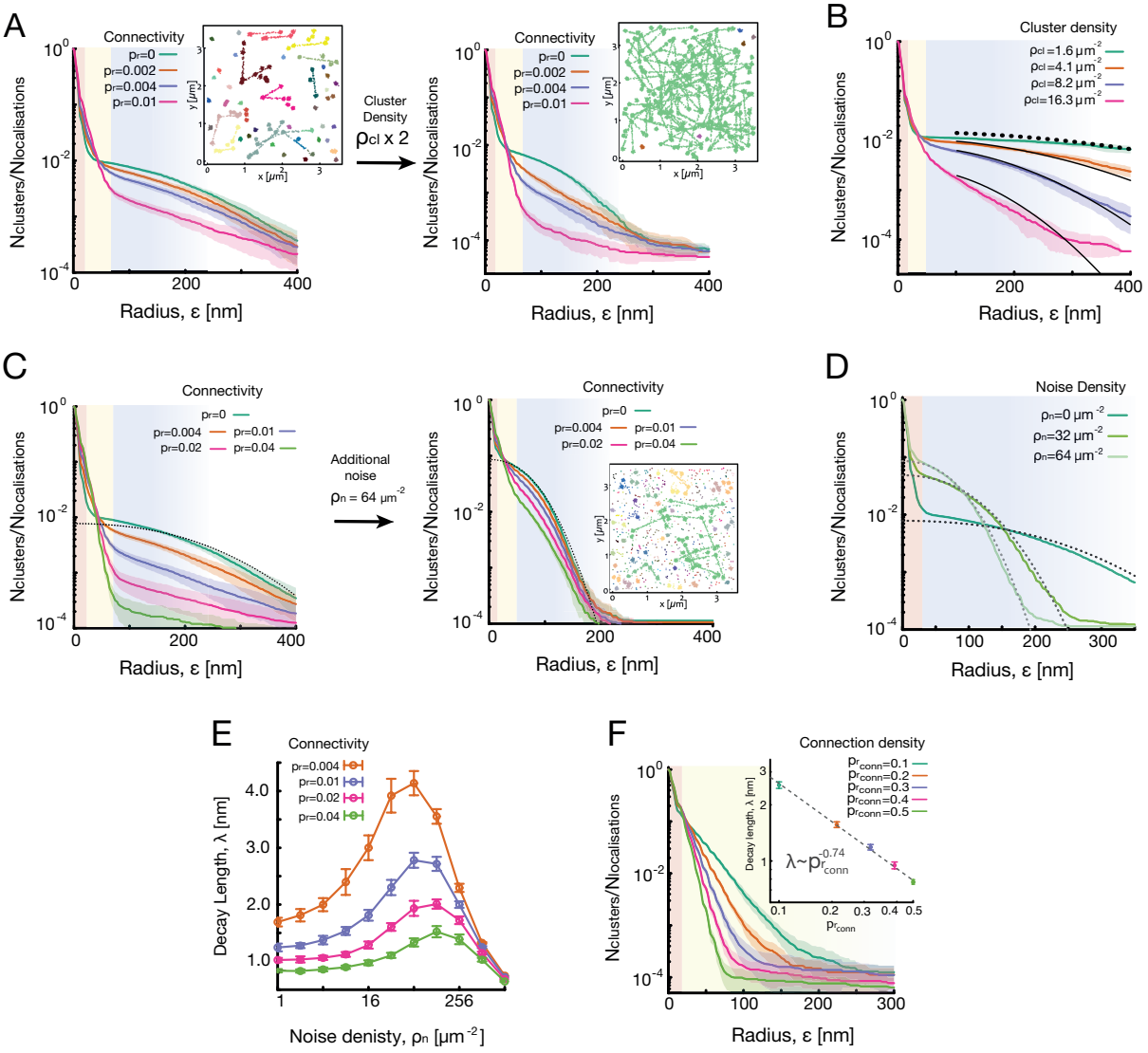

Figure S2. Average SuperStructure curves for different datasets. SuperStructure analysis was run on 20 independent datasets (in the same conditions) and the resulting curves were then averaged. Shaded regions represent the standard deviation from the average. Parameters are set to their standard values if not otherwise specified (see Methods). Palettes in the inset configurations represent cluster analysis at $\varepsilon = 80\ nm$. **A.** Locally connected clusters with different grades of connectivity and doubling the cluster density (from left to right): $\rho_{cl} = 8.2\ \mu m^{-2}$ (left) and $\rho_{cl} = 16.3\ \mu m^{-2}$ (right), connection density $p_{r_{conn}} = 0.5$, no noise and different values of connectivity p_r. The higher cluster density makes SuperStructure curves more markedly distinct as a function of p_r, compared to the same curves for a lower density. **B.** Locally connected clusters with low connectivity and increasing cluster density: connectivity $p_r = 0.002$, connection density $p_{r_{conn}} = 0.5$, no noise and different cluster densities ρ_{cl}. The first super-cluster regime maintains the single exponential decay, but the decay length λ decreases with the cluster density. In the main text, we showed that this dependence goes as $\lambda \propto \rho_{cl}^{-1/2}$. Also, the exponential decay λ_2 of the second super-cluster regime decreases with the density of clusters and this regime evolves from a Poisson-like (low ρ_{cl}) to an exponential decay (high ρ_{cl}). This behaviour seems to be a pure effect of the cluster density, as all other parameters remain unchanged. Black curve are Poisson decays attempts $\sim e^{-\pi\rho\varepsilon^2}$ to fit the second super-cluster regime. **C.** Locally connected clusters with different grades of connectivity and sparse noise addition: cluster density $\rho_{cl} = 8.2\ \mu m^{-2}$, connection density $p_{r_{conn}} = 0.5$, noise density $\rho_n = 0\ \mu m^{-2}$ (left) / $\rho_n = 64\ \mu m^{-2}$ (right) and different values of connectivity p_r. With high noise (8 times the cluster density), the 2nd super-cluster regime becomes Poissonian; the 1st super-cluster regime maintains its typical exponential decay, but the decay length is altered. Dotted lines represent fit with Eq.3 for $\varepsilon \in [70:300]\ nm$. **D.** Unconnected clusters of points with

increasing density of noise (other parameters are the same as **C.**). Eq.3 well describes the decay of the curves in the inter-cluster regime, with the density parameter ρ_{cl} and $\rho_{cl} + \rho_n$ respectively in absence and presence of noise. **E.** Average decay length of the first super-cluster regime for the connected systems represented in **C.** as function of noise density ρ_n. The fit to calculate the decay length λ has been made for $\varepsilon \in [20{,}60]\ nm$ for 20 independent datasets. Values of λ are then averaged. Bars represent the standard deviation from the average. Decay lengths for systems with different connectivities p_r are distinguishable as long as the noise density is below the connections density ($\sim 500\ \mu m^{-2}$). However, low noise density also alters the estimation of the decay length. The alteration is less severe for highly connected clusters. **F.** Fully connected meshes of clusters with increasing density of the mesh: cluster density $\rho_{cl} = 8.2\ \mu m^{-2}$, connectivity $p = 0.025$, no noise and different values of connection density $p_{r_{conn}}$. The super-clusters regime is unique, the decay is exponential and the decay length λ decreases with the density of the mesh. Fit of λ was performed for $\varepsilon \in [20{:}\,60]\ nm$. The inset shows the dependence of λ on $p_{r_{conn}}$ in a fully connected mesh, which is $\lambda \sim p_{r_{conn}}^{-0.74}$.

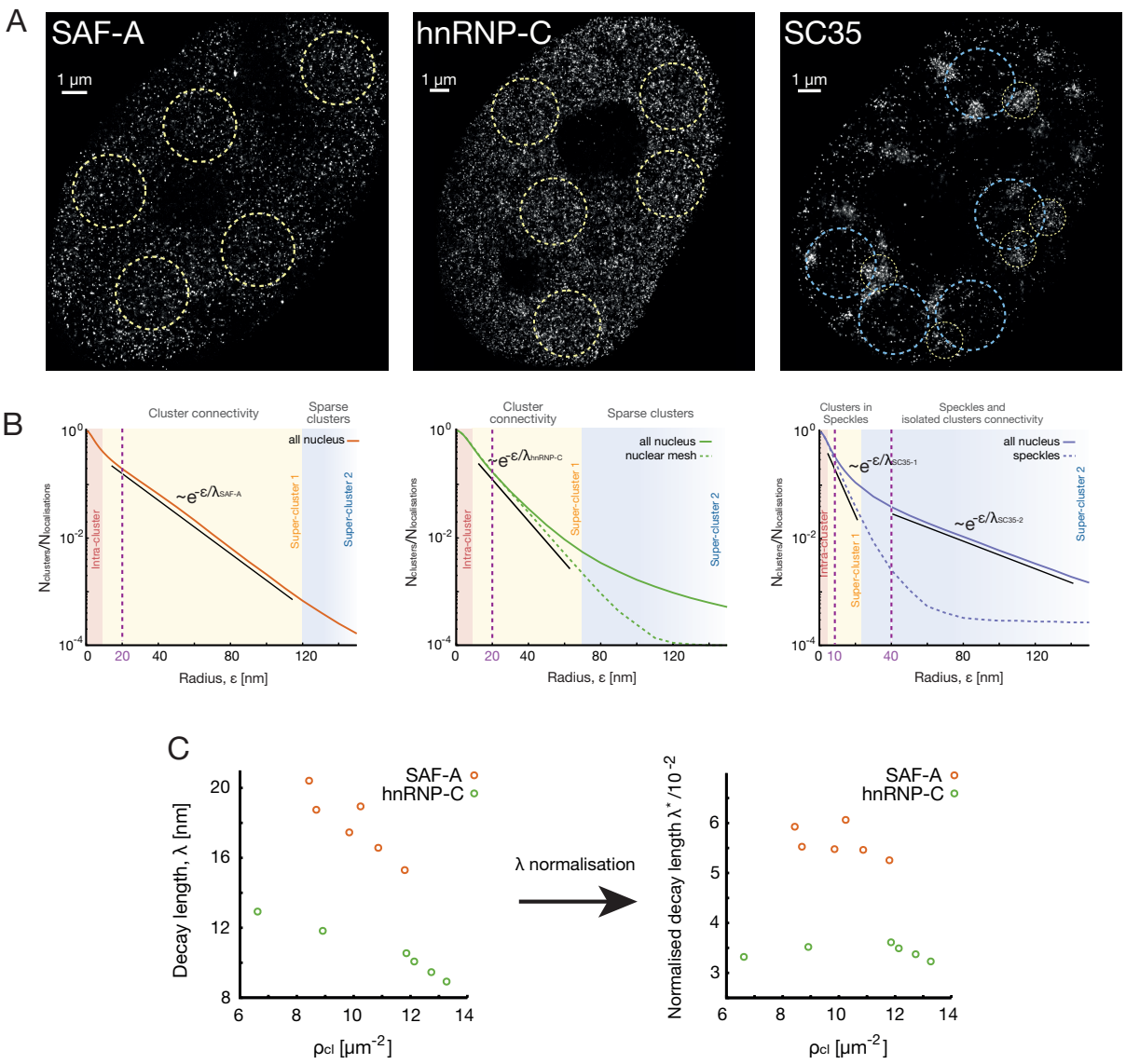

Figure S3. A. dSTORM reconstructed images of SAF-A, hnRNP-C and SC35 in a single cell where local circular regions for cluster density estimation purpose are highlighted. In case of SC35 two different region types are used, one inside speckles for the first exponential regime and one outside speckles for the second exponential regime. In the case of hnRNP-C and SC35 local circular regions were also used to compute SuperStructure "local curves" and the decay length λ in the first super-cluster regime as explained in Methods. **B.** Average SuperStructure curves for SAF-A, hnRNP-C and SC35 as shown and explained in the main text. Solid lines are the result of "all-nucleus' analysis, while dashed lines are the result of a "local" analysis (in local circular regions). Exponential regimes of interest are highlighted, as well as the values of ε at which the cluster analysis is made for clusters density estimation purpose (purple dashed vertical line). **C.** Check that connections are not the result of technical artefacts due to bad blinking quality both in SAF-A and hnRNP-C data by monitoring λ (left) and λ^* (right) for different cluster densities ρ_{cl}. The bad blinking quality of fluorophores would lead to localisation inaccuracy of emitters at the borders of protein clusters and in turn this could lead to pseudo-connections between clusters. However, these pseudo-connections would be proportional to the clusters density: higher cluster density would result in stronger pseudo-connections, which would reflect to a decrease of λ^* with the clusters density. λ, ρ_{cl} and λ^* were calculated for the 6 independent nuclei as explained in Methods and are shown in Table SI. Every nucleus can be considered as a system where the blinking conditions are the same, but clusters densities may vary due to statistical fluctuations. While, λ (left) decreases with ρ_{cl}, as expected, λ^* (right) is constant for different densities, ruling out the hypothesis that connections are artefacts due to bad blinking quality.

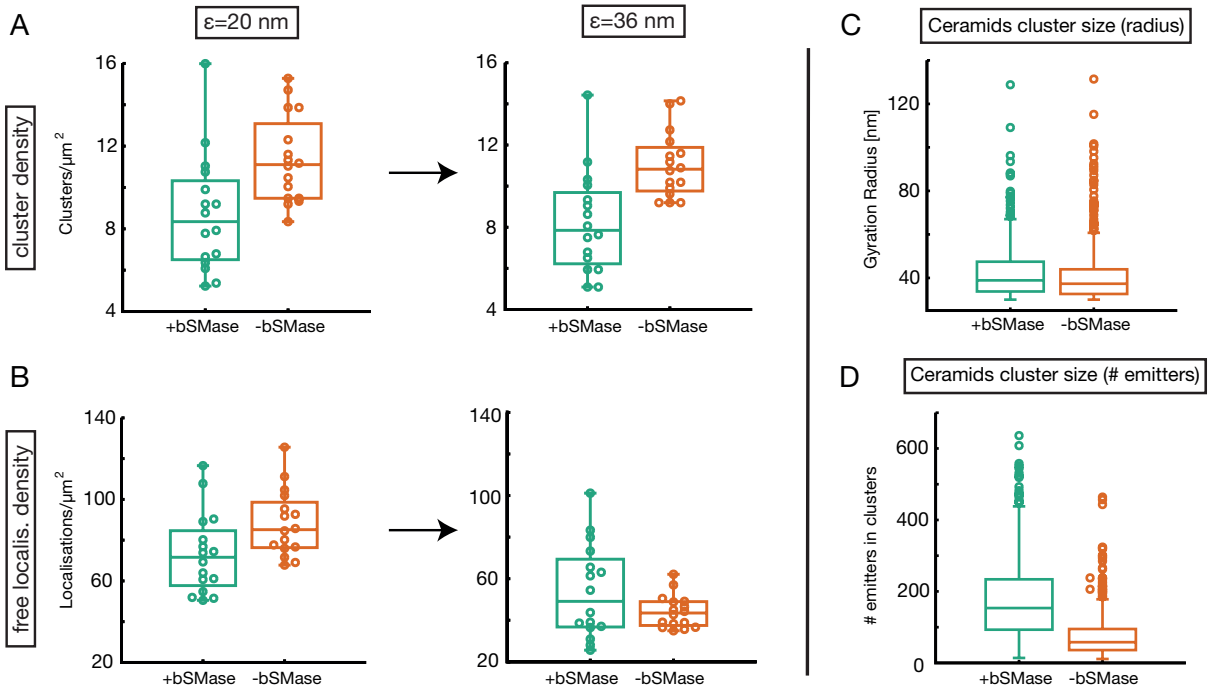

Figure S4. A. - B. The absence of local connectivity was confirmed by analysing cluster density (**A**) and sparse localisations density (**B**) in the cross-over range. We monitored the density of ceramides clusters and that of free emitters at $\varepsilon_1 = 20\,nm$ and $\varepsilon_2 = 36\,nm$. In order to calculate clusters density, DBSCAN was run at $N_{min} = 0$ and at the given value of ε and we kept only clusters with at least 10 particles. The remaining the particles were considered as free localisations. Clusters and free localisations were detected at $N_{min} = 0$ for 16 independent circular regions. The number of clusters remains constant in the considered ε regime, while the free localisations density significantly decreases, more severely for -bSMase cells. As a consequence, we can state that there is not significant merging of ceramides clusters, but only embedding of nearby free localisations in already formed clusters.
C. - D. Confirmation of the original paper results by calculating the ceramides cluster size both as gyration radius (**C**) and number of emitters (**D**). Protein clusters were detected at $N_{min} = 0$ at $\varepsilon^+ = 20\,nm$ and $\varepsilon^- = 24\,nm$. In accordance with the analysis in the paper, we looked at the size of clusters with a radius bigger than $30\,nm$. Note that +bSMase ceramides clusters consist (on average) in $180\,emitters$ in a circle of radius $42\,nm$. The resulting density is $32500\,\mu m^{-2}$. This result is approximately in line with our prediction obtained with the Poisson intra-cluster fit, by considering that the standard deviation of both cluster radius and emitters is high. Similarly, -bSMase clusters have on average $78\,emitters$ in an average cluster radius of $40\,nm$. The resulting density is $15500\,\mu m^{-2}$.

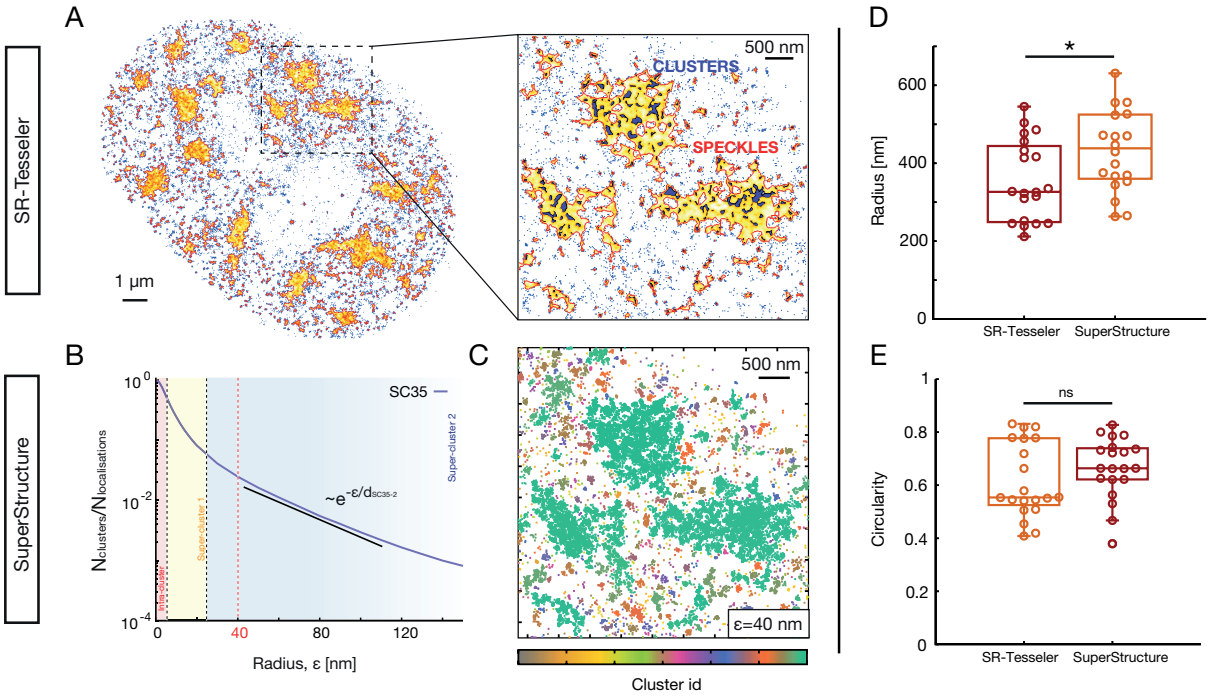

Figure S5. Size and shape estimation of local super-structures emerging in SC35 dSTORM data (i.e. nuclear speckles) by using both SuperStructure and SR-Tesseler. Analysis was performed on a single cell as proof of concept. **A.** Super-structures detection by using SR-Tesseler software, a segmentation framework based on Voronoï tessellation (constructed from the localisations coordinates). Adjustments of the density factor allows to detect structures at different density levels, such as clusters (violet) or speckles (yellow). Blue dots represent no-segmented localisations. The software was downloaded from *https://github.com/flevet/SR-Tesseler/releases/tag/v1.0* and run on a Windows OS. **B.** SuperStructure curve of the same data. Analysis of decay regimes allows to identify $\varepsilon = 40\ nm$ as a suitable value for super-structures identifications. **C.** Identified clusters at $\varepsilon = 40\ nm$ with SuperStructure. Speckles detections are visually compatible with those of SR-Tesseler. **D. - E.** Radius and circularity of super-structures by using both SR-Tesseler and SuperStructure. Both radius and circularity are very similar, showing the power of SuperStructure in computing shape and size properties. In the analysis we considered the 20 largest identified structures (i.e. speckles). SuperStructure: the 2d symmetric gyration tensor $\overrightarrow{R^2}$ was computed and diagonalised for identified super-structures. The gyration tensor components R^2_{xy} are defined as $R^2_{xy} = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j)$, where N is the total number of localisations in a super-structure, while x_i and y_i the x and y positions of the localisation i. The diagonalisation is necessary to obtain the major and minor axis of the speckles, namely γ_1 and γ_2. We then calculated the speckles radius $R_g = \sqrt{\gamma_1 + \gamma_2}$ and their circularity $c = \sqrt{\frac{|\gamma_1 - \gamma_2|}{\gamma_1 + \gamma_2}}$. SR-Tesseler: radius and circularity parameters were obtained as output after Voronoï tessellation. P-values were calculated using a Student's T-test: $ns\ P > 0.05$; $*\ P < 0.05$; $**\ P < 0.01$, $***\ P < 0.001$.

Supplemental Table

SAF-A			
Nucleus	$\lambda\,(nm)$	$\rho_{cl}(\mu m^{-2})$	$\lambda^*/10^{-2}$
1	18.75	8.686	5.526
2	17.46	9.846	5.477
3	18.94	10.24	6.062
4	15.30	11.80	5.255
5	16.57	10.87	5.463
6	20.41	8.432	5.926
Avg	17.90 ± 1.68	9.978 ± 1.173	5.618 ± 0.282

hnRNP-C			
Nucleus	$\lambda\,(nm)$	$\rho_{cl}(\mu m^{-2})$	$\lambda^*/10^{-2}$
1	11.82	8.912	3.520
2	12.93	6.621	3.320
3	10.07	12.13	3.492
4	9.463	12.27	3.374
5	8.920	13.27	3.229
6	10.54	11.86	3.613
Avg	10.62 ± 1.37	10.92 ± 2.37	3.425 ± 0.129

SC35-1 (first regime)			
Nucleus	$\lambda\,(nm)$	$\rho_{cl}(\mu m^{-2})$	$\lambda^*/10^{-2}$
1	5.882	23.17	2.693
2	5.094	32.85	2.898
3	4.777	36.92	2.818
4	4.797	38.96	2.976
5	4.591	35.65	2.534
6	7.033	19.10	2.937
Avg	5.362 ± 0.855	31.11 ± 7.38	2.809 ± 0.154

SC35-2 (second regime)			
Nucleus	$\lambda\,(nm)$	$\rho_{cl}(\mu m^{-2})$	$\lambda^*/10^{-2}$
1	36.02	5.517	8.461
2	29.46	4.838	6.479
3	27.14	5.404	6.309
4	35.16	4.527	7.481
5	31.48	4.584	6.740
6	30.33	4.951	6.748
Avg	31.60 ± 3.11	4.970 ± 0.377	7.036 ± 0.735

Table SI. Decay length λ, detected clusters density ρ_{cl} and normalised decay length $\lambda^* = \lambda/\rho_{cl}^{-1/2}$ for SAF-A, hnRNP-C and SC-35 (in both super-cluster regimes SC35-1 and SC35-2). Both single-nucleus values and average over nuclei ($\pm$ standard deviation) are shown. For SAF-A and SC35-2, λ was obtained by fitting "all-nucleus" SuperStructure curves, i.e. curves where the entire nucleus was analysed. On the other hand, for hnRNP-C and SC35-1, λ was obtained by fitting "local" SuperStructure curves, i.e. curves where local circular regions were analysed as explained in Methods. In the latter case, nuclear values showed in the table are the result of an average over 5 independent "local" values within the same cell.

References

Baumgart, F., A.M. Arnold, K. Leskovar, K. Staszek, M. Fölser, J. Weghuber, H. Stockinger, and G.J. Schütz. 2016. Varying label density allows artifact-free analysis of membrane-protein nanoclusters. *Nat. Methods*. 13:661–664. doi:10.1038/nmeth.3897.

Beliveau, B.J., A.N. Boettiger, M.S. Avendaño, R. Jungmann, R.B. McCole, E.F. Joyce, C. Kim-Kiselak, F. Bantignies, C.Y. Fonseka, J. Erceg, M.A. Hannan, H.G. Hoang, D. Colognori, J.T. Lee, W.M. Shih, P. Yin, X. Zhuang, and C.T. Wu. 2015. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat. Commun.* 6. doi:10.1038/ncomms8147.

Bintu, B., L.J. Mateo, J.H. Su, N.A. Sinnott-Armstrong, M. Parker, S. Kinrot, K. Yamaya, A.N. Boettiger, and X. Zhuang. 2018. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science (80-.).* 362.

Boettiger, A.N., B. Bintu, J.R. Moffitt, S. Wang, B.J. Beliveau, G. Fudenberg, M. Imakaev, L.A. Mirny, C. Wu, and X. Zhuang. 2016. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*. 529:418–422. doi:10.1038/nature16496.

Brangwynne, C.P., P. Tompa, and R. V Pappu. 2015. Polymer physics of intracellular phase transitions. *Nat. Phys.* 11:899–904. doi:10.1038/nphys3532.

Bronshtein, I., E. Kepten, I. Kanter, S. Berezin, M. Lindner, A.B. Redwood, S. Mai, S. Gonzalo, R. Foisner, Y. Shav-Tal, and Y. Garini. 2015. Loss of lamin A function increases chromatin dynamics in the nuclear interior. *Nat. Commun.* 6:1–9. doi:10.1038/ncomms9044.

Burgert, A., J. Schlegel, J. Bécam, S. Doose, E. Bieberich, A. Schubert-Unkmeir, and M. Sauer. 2017. Characterization of Plasma Membrane Ceramides by Super-Resolution Microscopy. *Angew. Chemie*. 129:6227–6231.

Caudron-Herger, M., K. Müller-Ott, J.P. Mallm, C. Marth, U. Schmidt, K. Fejes-Tóth, and K. Rippe. 2011. Coding RNAs with a non-coding function: Maintenance of open chromatin structure. *Nucleus*. 2.

Cho, W.-K., J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I.I. Cisse. 2018. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science (80-.).* 361:412–415. doi:10.1126/science.aar4199.

Chong, S., C. Dugast-Darzacq, Z. Liu, P. Dong, G.M. Dailey, C. Cattoglio, A. Heckert, S. Banala, L. Lavis, X. Darzacq, and R. Tjian. 2018. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science (80-.).* 361.

Cisse, I.I., I. Izeddin, S.Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, and X. Darzacq. 2013. Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. *Science (80-.).* 341:664–667. doi:10.1126/science.1239053.

Deniz, A.A., S. Mukhopadhyay, and E.A. Lemke. 2008. Single-molecule biophysics: At the interface of biology, physics and chemistry. *J. R. Soc. Interface*. 5:15–45. doi:10.1098/rsif.2007.1021.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press. 226–231.

Frank, L., and K. Rippe. 2020. Repetitive RNAs as Regulators of Chromatin-Associated Subcompartment Formation by Phase Separation. *J. Mol. Biol.* 432:4270–4286.

Garcia-Parajo, M.F., A. Cambi, J.A. Torreno-Pina, N. Thompson, and K. Jacobson. 2014. Nanoclustering as a dominant feature of plasma membrane organization. *J. Cell Sci.* 127:4995–5005.

Goundaroulis, D., E. Lieberman Aiden, and A. Stasiak. 2019. Chromatin Is Frequently Unknotted at the Megabase Scale. *Biophys. J.* 118:2268–2279. doi:10.1016/j.bpj.2019.11.002.

Griffié, J., M. Shannon, C.L. Bromley, L. Boelen, G.L. Burn, D.J. Williamson, N.A. Heard, A.P. Cope, D.M. Owen, and P. Rubin-Delanchy. 2016. A Bayesian cluster analysis method for single-molecule localization microscopy data. *Nat. Protoc.* 11:2499–2514. doi:10.1038/nprot.2016.149.

Hall, L.L., and J.B. Lawrence. 2016. RNA as a fundamental component of interphase chromosomes: Could repeats prove key? *Curr. Opin. Genet. Dev.* 37:137–147. doi:10.1016/j.gde.2016.04.005.

Hennig, S., S. van de Linde, S. Bergmann, T. Huser, and M. Sauer. 2015. Quantitative Super-Resolution Microscopy of Nanopipette-Deposited Fluorescent Patterns. *ACS Nano*. 9:8122–8130.

Henriques, R., C. Griffiths, E.H. Rego, and M.M. Mhlanga. 2011. PALM and STORM: Unlocking live-cell super-resolution. *Biopolymers*. 95:322–331. doi:10.1002/bip.21586.

Huang, B., W. Wang, M. Bates, and X. Zhuang. 2008. Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy. *Science (80-.).* 319:810–813. doi:10.1126/science.1153529.

Izeddin, I., C.G. Specht, M. Lelek, X. Darzacq, A. Triller, C. Zimmer, and M. Dahan. 2011. Super-Resolution Dynamic Imaging of Dendritic Spines Using a Low-Affinity Photoconvertible Actin Probe. *PLoS One*. 6:1–14.

Jackson, D.A., A. Pombo, and F. Iborra. 2000. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *FASEB J.* 14:242–254.

Kapanidis, A.N., A. Lepore, and M. El Karoui. 2018. Rediscovering Bacteria through Single-Molecule Imaging in Living Cells. *Biophys. J.* 115:190–202. doi:10.1016/j.bpj.2018.03.028.

Khanna, N., Y. Zhang, J.S. Lucas, O.K. Dudko, and C. Murre. 2019. Chromosome dynamics near the sol-gel phase transition dictate the timing of remote genomic interactions. *Nat. Commun.* 10:1–13. doi:10.1038/s41467-019-10628-9.

Larson, A.G., D. Elnatan, M.M. Keenen, M.J. Trnka, J.B. Johnston, A.L. Burlingame, D.A. Agard, S. Redding, and G.J. Narlikar. 2017. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*. 547:236–240. doi:10.1038/nature22822.

Leidescher, S., J. Nuebler, Y. Feodorova, E. Hildebrand, S. Ullrich, S. Bultmann, S. Link, K. Thanisch, J. Dekker, H. Leonhardt, L. Mirny, and I. Solovei. 2020. Spatial Organization of Transcribed Eukaryotic Genes. *bioRxiv*. 2020.05.20.106591. doi:10.1101/2020.05.20.106591 (Preprint posted May 21, 2020).

Levet, F., E. Hosy, A. Kechkar, C. Butler, A. Beghin, D. Choquet, and J.-B. Sibarita. 2015. SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. *Nat. Methods*. 12:1065–1071. doi:10.1038/nmeth.3579.

Lin, S., G. Coutinho-Mansfield, D. Wang, S. Pandit, and X.-D. Fu. 2008. The splicing factor SC35 has an active role in transcriptional elongation. *Nat. Struct. Mol. Biol.* 15:819–826. doi:10.1038/nsmb.1461.

van de Linde, S., A. Löschberger, T. Klein, M. Heidbreder, S. Wolter, M. Heilemann, and M. Sauer. 2011. Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat. Protoc.* 6:991–1009. doi:10.1038/nprot.2011.336.

van de Linde, S., and M. Sauer. 2014. How to switch a fluorophore: from undesired blinking to controlled photoswitching. *Chem. Soc. Rev.* 43:1076–1087.

Maharana, S., J. Wang, D.K. Papadopoulos, D. Richter, A. Pozniakovsky, I. Poser, M. Bickle, S. Rizk, J. Guillén-Boixet, T.M. Franzmann, M. Jahnel, L. Marrone, Y.T. Chang, J. Sterneckert, P. Tomancak, A.A. Hyman, and S. Alberti. 2018. RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science (80-.).* 360:918–921. doi:10.1126/science.aar7366.

Maiser, A., S. Dillinger, G. Längst, L. Schermelleh, H. Leonhardt, and A. Németh. 2020. Super-resolution in situ analysis of active ribosomal DNA chromatin organization in the nucleolus. *Sci. Rep.* 10:7462. doi:10.1038/s41598-020-64589-x.

Malkusch, S., and M. Heilemann. 2016. Extracting quantitative information from single-molecule super-resolution imaging data with LAMA – LocAlization Microscopy Analyzer. *Sci. Rep.* 6:34486. doi:10.1038/srep34486.

McSwiggen, D.T., A.S. Hansen, S.S. Teves, H. Marie-Nelly, Y. Hao, A.B. Heckert, K.K. Umemoto, C. Dugast-Darzacq, R. Tjian, and X. Darzacq. 2019. Evidence for DNA-mediated nuclear compartmentalization distinct from phase separation. *Elife*. 8:e47098. doi:10.7554/eLife.47098.

Michieletto, D., and N. Gilbert. 2019. Role of nuclear RNA in regulating chromatin structure and transcription. *Curr. Opin. Cell Biol.* 58:120–125. doi:10.1016/j.ceb.2019.03.007.

Nir, G., I. Farabella, C. Pérez Estrada, C.G. Ebeling, B.J. Beliveau, H.M. Sasaki, S.H. Lee, S.C. Nguyen, R.B. McCole, S. Chattoraj, J. Erceg, J. AlHaj Abed, N.M.C. Martins, H.Q. Nguyen, M.A. Hannan, S. Russell, N.C. Durand, S.S.P. Rao, J.Y. Kishi, P. Soler-Vila, M. Di Pierro, J.N. Onuchic, S.P. Callahan, J.M. Schreiner, J.A. Stuckey, P. Yin, E.L. Aiden, M.A. Marti-Renom, and C.T. Wu. 2018. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.* 14:1–35. doi:10.1371/journal.pgen.1007872.

Nozawa, R.-S., L. Boteva, D.C. Soares, C. Naughton, A.R. Dun, A. Buckle, B. Ramsahoye, P.C. Bruton, R.S. Saleeb, M. Arnedo, B. Hill, R.R. Duncan, S.K. Maciver, and N. Gilbert. 2017. SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs. *Cell*. 169:1214-1227.e18. doi:10.1016/j.cell.2017.05.029.

Ovesný, M., P. Křížek, J. Borkovec, Z. Švindrych, and G.M. Hagen. 2014. ThunderSTORM: A comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics*. 30:2389–2390.

Owen, D.M., C. Rentero, J. Rossy, A. Magenau, D. Williamson, M. Rodriguez, and K. Gaus. 2010. PALM imaging and cluster analysis of protein heterogeneity at the cell surface. *J. Biophotonics*. 3:446–454. doi:10.1002/jbio.200900089.

Prakash, K., D. Fournier, S. Redl, G. Best, M. Borsos, V.K. Tiwari, K. Tachibana-Konwalski, R.F. Ketting, S.H. Parekh, C. Cremer, and U.J. Birk. 2015. Superresolution imaging reveals structurally distinct periodic patterns of

chromatin along pachytene chromosomes. *Proc. Natl. Acad. Sci. United States Am.* 112:14635–14640. doi:10.1073/pnas.1516928112.
Resch, G.P., K.N. Goldie, A. Krebs, A. Hoenger, and J.V. Small. 2002. Visualisation of the actin cytoskeleton by cryo-electron microscopy. *J. Cell Sci.* 115:1877–1882.
Revyakin, A., C. Liu, R.H. Ebright, and T.R. Strick. 2006. Abortive Initiation and Productive Initiation by RNA Polymerase Involve DNA Scrunching Andrey. *Science (80-.).* 314:1139–1144.
Rogers, S.L., U. Wiedemann, N. Stuurman, and R.D. Vale. 2003. Molecular requirements for actin-based lamella formation in Drosophila S2 cells. *J. Cell Biol.* 162:1079–1088.
Sabari, B.R., A. Dall'Agnese, A. Boija, I.A. Klein, E.L. Coffey, K. Shrinivas, B.J. Abraham, N.M. Hannett, A. V Zamudio, J.C. Manteiga, C.H. Li, Y.E. Guo, D.S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T.I. Lee, I.I. Cisse, R.G. Roeder, P.A. Sharp, A.K. Chakraborty, and R.A. Young. 2018. Coactivator condensation at super-enhancers links phase separation and gene control. *Science (80-.).* 361.
Sauer, M., and M. Heilemann. 2017. Single-Molecule Localization Microscopy in Eukaryotes. *Chem. Rev.* 117:7478–7509.
Schermelleh, L., R. Heintzmann, and H. Leonhardt. 2010. A guide to super-resolution fluorescence microscopy. *J. Cell Biol.* 190:165–175.
Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D.J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. 2012. Fiji: an open-source platform for biological-image analysis. *Nat. Methods*. 9:676–682. doi:10.1038/nmeth.2019.
Sengupta, P., T. Jovanovic-Talisman, D. Skoko, M. Renz, S.L. Veatch, and J. Lippincott-Schwartz. 2011. Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. *Nat. Methods*. 8:969–975. doi:10.1038/nmeth.1704.
Siegberg, D., and D.-P. Herten. 2011. Fluorescence Quenching of Quantum Dots by DNA Nucleotides and Amino Acids1. *Aust. J. Chem.* 64:512–516.
Smeets, D., Y. Markaki, V.J. Schmid, F. Kraus, A. Tattermusch, A. Cerase, M. Sterr, S. Fiedler, J. Demmerle, J. Popken, H. Leonhardt, N. Brockdorff, T. Cremer, L. Schermelleh, and M. Cremer. 2014. Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. *Epigenetics Chromatin*. 7:8. doi:10.1186/1756-8935-7-8.
Spahn, C., F. Herrmannsdörfer, T. Kuner, and M. Heilemann. 2016. Temporal accumulation analysis provides simplified artifact-free analysis of membrane-protein nanoclusters. *Nat. Methods*. 13:963–964. doi:10.1038/nmeth.4065.
Strom, A.R., A. V Emelyanov, M. Mir, D. V Fyodorov, X. Darzacq, and G.H. Karpen. 2017. Phase separation drives heterochromatin domain formation. *Nature*. 547:241–245. doi:10.1038/nature22989.
Szabo, Q., D. Jost, J.M. Chang, D.I. Cattoni, G.L. Papadopoulos, B. Bonev, T. Sexton, J. Gurgo, C. Jacquier, M. Nollmann, F. Bantignies, and G. Cavalli. 2018. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Sci. Adv.* 4:1–14. doi:10.1126/sciadv.aar8082.
Wang, S., J. Su, B.J. Beliveau, B. Bintu, J.R. Moffitt, and C. Wu. 2016. Spatial organization of chromatin domains and compartments in single chromosomes.

Science (80-.). 353:598.
Williamson, D.J., G.L. Burn, S. Simoncelli, J. Griffié, R. Peters, D.M. Davis, and D.M. Owen. 2020. Machine learning for cluster analysis of localization microscopy data. *Nat. Commun.* 11:1493. doi:10.1038/s41467-020-15293-x.
Xiao, R., P. Tang, B. Yang, J. Huang, Y. Zhou, C. Shao, H. Li, H. Sun, Y. Zhang, and X.-D. Fu. 2012. Nuclear Matrix Factor hnRNP U/SAF-A Exerts a Global Control of Alternative Splicing by Regulating U2 snRNP Maturation Rui. *Mol. Cell*. 45:656–668. doi:10.1038/jid.2014.371.
Xie, L., P. Dong, X. Chen, T.-H.S. Hsieh, S. Banala, M. De Marzio, B.P. English, Y. Qi, S.K. Jung, K.-R. Kieffer-Kwon, W.R. Legant, A.S. Hansen, A. Schulmann, R. Casellas, B. Zhang, E. Betzig, L.D. Lavis, H.Y. Chang, R. Tjian, and Z. Liu. 2020. 3D ATAC-PALM: super-resolution imaging of the accessible genome. *Nat. Methods*. 17:430–436. doi:10.1038/s41592-020-0775-2.
Xie, S.Q., S. Martin, P. V Guillot, D.L. Bentley, and A. Pombo. 2006. Splicing Speckles Are Not Reservoirs of RNA Polymerase II, but Contain an Inactive Form, Phosphorylated on Serine2 Residues of the C-Terminal Domain. *Mol. Biol. Cell*. 17:1723–1733. doi:10.1091/mbc.e05-08-0726.